2012

Essays on statistical inference with imperfectly observed data

Hang Qian
Iowa State University

Essays on statistical inference with imperfectly observed data

by

Hang Qian

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Economics

Program of Study Committee:

Brent Kreider, Major Professor

John Schroeter

Joseph Herriges

Oleksandr Zhylyevskyy

Petrutza Caragea

Iowa State University

Ames, Iowa

2012

DEDICATION

I would like to dedicate this dissertation to my parents without whose support I would not have been able to complete this work. I can hardly find the words to express my gratitude for your unfailing love and support along this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am greatly indebted to those who helped me with various aspects of conducting research and writing papers. First and foremost, I would like to take this opportunity to express my thanks to Dr. Brent Kreider for his guidance, patience and support throughout my research and the writing of this dissertation. His insights and words of encouragement have often inspired me towards new height of research. I would also like to thank my dissertation committee members for their efforts and contributions to this work: Dr. John Schroeter, Dr. Joseph Herriges, Dr. Oleksandr Zhylyevskyy and Dr. Petrutza Caragea. I would additionally like to thank Dr. Rajesh Singh and Dr. Helle Bunzel for their invaluable help in my research.

ABSTRACT

Missing data is a common problem encountered by empirical researchers and practitioners. This dissertation is a collection of three essays on handling imperfectly observed economic data. The first essay addresses temporal aggregation where some high frequency data are missing but their sum or average are observed in the form of low frequency data. In a vector autoregression model with varied frequency data, the explicit form of the likelihood function and the posterior distribution of missing values are found without resorting to the recursive Kalman filter. The second essay further discusses data aggregation in a two-equation model in which the missing values are imputed by a regression. In two scenarios, the likelihood function is shown to be separable and the analytic maximum likelihood estimator can be obtained by two auxiliary regressions, which is advantageous to the conventional least squares imputation approach in terms of both efficiency and computability. The third essay concerns the finite-sample bias of estimators associated with the monotone instrumental variables, which is a useful assumption to partially identify the counterfactual outcomes. It is shown that a multi-level bootstrap procedure can reduce and gradually eliminate the bias. A simultaneous simulation strategy is also proposed to make multi-level bootstrap computationally feasible.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Topics of the dissertation

Most economic data, at the microeconomic or macroeconomic level alike, are observational. A researcher does not have full control on the quality of the data collected for empirical studies. Sooner or later, one will encounter the missing data problem. Missing data is a broad concept that encompasses many possibilities. First, sometimes a few observations are missing at random, which is relatively easy to handle. Two quick solutions are either to discard the missing values in the regression or interpolate them by *ad hoc* mathematical procedures such as polynomial fillings. More advanced treatments include multiple imputation, expectation-maximization and Bayesian data augmentation methods. Second, sometimes disaggregated data are unobservable to the researcher but their sum or average can be observed. This is the data aggregation problem. Though one can align the variables either by aggregating the observables or interpolating the disaggregated missing data, the problem can be more effectively handled by exploring the underlying data generating process and the aggregation constraints. Third, there are cases in which some non-experimental data are counterfactual and can never be observed. The huge body of literature on the treatment effect identification attempts to infer the counterfactuals from the observables under proper identification assumptions.

This dissertation is a collection of three essays on handling imperfectly observed economic data. The first essay addresses the temporal aggregation. It is motivated by the fact that macroeconomic data are not observed at a uniformed frequency and the best available data could be, for example, a monthly-quarterly mixture. A tractable approach that makes full use of data at varied frequencies is proposed in a Bayesian framework. It contributes to the literature by articulating the closed-form likelihood and posterior latent variables without resorting

to the recursive formula prescribed by the Kalman filter. The second essay revisits an old two-equation regression model on an aggregation problem mostly suitable for the microeconomic data analysis. The first equation regresses the dependent variable of interest on a set of covariates. However, one key covariate does not have disaggregated data, which are imputed by the second regression equation. The main contribution of that paper is the discovery of an analytic maximum likelihood estimator obtained by two auxiliary regressions. That implies an efficient estimator can be obtained without computational barriers. The third essay discusses finite-sample bias correction of an estimator associated with the monotone instrumental variables, which is a useful assumption for identifying treatment effects. The innovative part of the paper is a multi-level bootstrap procedure, which is shown to effectively reduce the finite-sample bias. In addition, higher level bootstrap does not necessarily suffer from the curse of dimensionality, since a simultaneous simulation strategy can make multi-level bootstrap computationally feasible.

1.2 Organization of the dissertation

The rest of the dissertation will be organized as follows.

Chapter 2 discusses temporal aggregation in the context of a vector autoregression model with varied frequency data. After setting up the model that allows data being observed at arbitrarily mixed frequencies, we make explicit the likelihood function and the posterior latent variables by exploring the covariance structure of the time series and the aggregation constraints. Then the new approached is compared with the conventional Kalman filter solutions. The major difference is that our method uses all the information as a whole while the Kalman filter updates and assimilates information date by date. The paper also considers using information block by block. The block Gibbs sampler is inspired by the Markov property of the autoregressive series. The sampler runs fast on evenly aggregated data. Lastly, the approach is applied to a structural vector autoregression model that describes dynamic effects of the monetary policy on future outputs and prices. Using varied frequency data effectively relaxes the identification assumption that monetary policy shocks have no contemporaneous effects on the real economy.

Chapter 3 further studies the data aggregation problem in a two-equation regression model. We revisit the model and find that the likelihood function is separable by suitable reparameterization if one instrument corresponds to one endogenous regressor. In that case, an analytic full-information maximum likelihood estimator exists and can be obtained by two auxiliary regressions. For a comparison with our maximum likelihood estimator, the properties of least square solutions to the model are also discussed. For regressions with endogeneity problems, some LS estimators are not consistent, and some consistent estimators discard apparent information. Those drawbacks are overcome by the ML estimator, which is advantageous in terms of both efficiency and computability. Chapter 3 also discusses a Bayesian solution to the two-equation model with aggregated data. The proposed Gibbs sampler is most useful in flexible settings when the likelihood function does not satisfy the separability condition. For models without analytic solutions, Monte Carlo studies show that the Bayesian estimator is more robust and less sensitive to the initial values.

Chapter 4 addresses the finite sample bias induced by the monotone instrumental variables models, which feature a supremum operator in the lower bound and an infimum operator in the upper bound. However, when sampling variation is taken into account, the analogue estimate of the lower bound is biased upwards and upper bound biased downwards, resulting in estimates that are narrower than the true bounds. We first propose a conservative estimator which is biased in the opposite but more favorable direction. Then under a polynomial approximation assumption, we show the mechanism of the parametric bootstrap correction procedure, which can reduce but not eliminate the bias with a possibility of overcorrection. The inadequacy of the single bootstrap motivates us to pursue higher level bootstraps, which are shown to be able to further reduce the bias. Furthermore, a simultaneous simulation strategy can be used to make multi-level bootstraps computationally feasible. Lastly, we apply our estimators to a disability misreporting problem in health economics. Both the conservative estimator and the multi-level bootstrap corrected estimator work well.

Chapter 5 concludes the dissertation by explaining the limitations of the proposed approaches and suggesting some directions for future research.

CHAPTER 2. VECTOR AUTOREGRESSION WITH VARIED FREQUENCY DATA

2.1 Introduction

The Vector Autoregression (VAR) proposed by Christopher Sims (1980) is a workhorse model for forecasting as well as studying cause and effect in the macroeconomy. An autoregression model implicitly assumes data are sampled at the same frequency since variables at date t are regressed on variables dated at $t-1, t-2$, etc. However, macroeconomic data are not always observed at a uniformed frequency. First, each series can be sampled at its own frequency. For example, the best available data of GDP is quarterly, while that of the CPI is monthly, that of financial asset returns might be daily or more frequent. Second, for a given variable, recent data may be observed at a higher frequency while historical data are coarsely sampled. For instance, quarterly GDP data are not available until 1947.

In the presence of varied frequency data, a VAR practitioner usually aligns variables either downward by aggregating the data to a lower frequency or upward by interpolating the high frequency data with heuristic rules such as polynomial fillings. Downward alignment discards valuable information in the high frequency data. Furthermore, temporal aggregation can change the lag order of ARMA models (Amemiya and Wu, 1972), reduce efficiency in parameter estimation and forecast (Tiao and Wei, 1976), affect Granger-causality and cointegration among component variables (Marcellino, 1999), induce spurious instantaneous causality (Breitung and Swanson, 2002), and so on. Silvestrini and Veredas (2008) provide a comprehensive review on the theory of temporal aggregation. On the other hand, upward alignment on the basis of *ad hoc* mathematical procedures is also problematic. Pavia-Miralles (2010) surveys various methods of interpolating and extrapolating time series. The problem is that by using a VAR model

we acknowledge high frequency data are generated by that model. However, the interpolation is not based on the multivariate model that generates the data, but on other heuristic rules, which inevitably introduces noises, if not distortion, to the data.

Frequency mismatch is essentially a missing data problem in which some high frequency data are unobservable. An effective way to handle missing data in linear time series models is to use the state space representation and apply the Kalman filter. Jones (1980) pioneers this approach by writing an ARMA model in the state space form, skipping the missing values by setting the updated states equal to one-period predicted states. The Kalman filter effectively marginalizes the missing data out of the likelihood function and obtains the likelihood of observed data in its prediction error decomposition form.

There are other methods handling missing data in time series. One sensible approach is to fill the missing data with an arbitrary value and meanwhile include an additive outliers dummy. Gomez et al. (1998) and Proietti (2008) show the connections between the outliers approach and the Kalman filter solution. Another approach is to treat (at least superficially) the missing values as if they were unknown parameters. Some variants of expectation-maximization algorithm may be adopted for parameter estimation. See Sargan and Drettakis (1974), Pena and Tiao (1991), Stoica et al. (2005) for discussions. The third approach is based on moments derived from the Yule-Walker equations (Chen and Zadrozny, 1998). In the presence of systematic missing data not all moments are estimable, but some computable analogue moments can be used to estimate parameters in a generalized method of moments framework.

In this paper, we discuss the estimation strategy of a special type of missing data problem, namely temporal aggregation. Any scenario of high frequency data lost can be called missing data, but temporal aggregation also features observation on the sum or average of these lost high frequency data. For example, flow variables such as quarterly GDP is the sum of the latent "monthly GDP" in a quarter. Stock variables such as the monthly CPI are more reasonably viewed as an average of the latent "weekly CPI" in a month instead of the price level in the last week of a month.[1]

[1]in the literature, stock variables are defined as those sampled every k periods from the latent high frequency variables. The examples provided for stock variables are "rates and indexes" such as interest rate, unemployment rate and CPI (Silvestrini and Veredas, 2008). However, all of them seem to be generated by averaging the latent

In the literature, the only available method to handle temporal aggregation is through the Kalman filter. The seminal paper of Harvey and Pierse (1984) outlines the state space representation of an ARMA model subject to temporal aggregation. The idea is to enlarge the state vector by including recent disaggregated (high frequency) variables, so that the observed aggregated (low frequency) variables can be expressed as the sum of latent states. The mixed frequency VAR and a related factor model have been explored by Zadrozny (1988), Mittnik and Zadrozny (2004), Mariano and Murasawa (2003, 2010), Hyung and Granger (2008), all of which rely on the state space representation. Also note that other approaches addressing missing data problems, such as additive outliers, missing data as parameters and Yule-Walker moments, cannot be applied to the temporal aggregation problems, because the sum or average of these lost high frequency data are also informative.

We contribute to the literature by handling the temporal aggregation problem from a new perspective. The mixed frequency data provide us two pieces of information. The first is partially observed high frequency data. The second is some low frequency data. On the one hand, by employing an VAR model we agree that disaggregated data are generated by such model. In other words, the disaggregated variables have a joint distribution conformable with the autocorrelation structure of the VAR model. Therefore, we can find out the distribution of missing high frequency data conditional on observed high frequency data. On the other hand, the observed low frequency data impose linear constraints on the distribution of the missing high frequency variables. Combining the two pieces of information, we can obtain an explicit solution to both the likelihood function of observed data and the posterior latent data, without resorting to an recursive formula offered by the Kalman filter.

The difference between our approach and the Kalman filter can be intuitively expressed as follows. Every period there are some informative high and/or low frequency data. The Kalman filter assimilates information date by date and consecutively updates distributions of latent states as new information arrives. In many engineering applications of the Kalman filter such as real-time controls, the recursiveness feature offers a great advantage in that only increments

variables in the k periods. Some financial data such as the S&P 500 index and exchange rates in a week or month do have their last-trading-day data available, but the averaged data are offered at the same time.

of information, rather than the whole information set, are used for predicting and updating current latent states. However, in statistics and economics applications, the dataset containing historical observations is usually fixed in length. If all information is readily available, why not use information of all dates together? Our approach focuses on the joint distribution of multiple-period disaggregated variables, which are bound by aggregated observations.

This difference also carries empirical implications. Our approach addresses data aggregation in an explicit and straightforward manner. The setup of the model is general enough to allow linear temporal aggregation of any types. Though for a given aggregation structure it is always possible to design its state space representation, but the design must be tailored and finished by the practitioners. However, our method is more friendly to users, since it only requires users to provide the data (say, in a spreadsheet), while the estimation is as routine as a standard VAR model.

Though our approach articulates both the likelihood and posteriors in an explicit form, we prefer to put the estimation in a Bayesian framework. It is easier to formulate the likelihood function than to maximize it, for a VAR model typically contains many coefficients and numerical algorithms such as the quasi-Newton have limited ability to implement the maximum likelihood estimation.[2] The advantage of adopting the Bayesian framework is that the Gibbs sampler disentangles two distinct tasks: i) estimating model parameters conditional on complete data; ii) decoding latent variables conditional on model parameters. The large number of parameters pose little computational challenges in that they are handled in a linear regression model if complete data were observed. As for decoding latent variables, recently Viefers (2011) uses the Kalman filter to sample the smoothed disaggregated variables, while we articulate the posterior conditional distribution of disaggregated variables explicitly as a multivariate normal distribution subject to several linear constraints.

Our idea of sampling latent variables is somewhat close to a Bayesian estimation of the VAR

[2]As is noted by Chen and Zadrozny (1998), Kalman filter method may perform poorly or not at all on a larger model. In applications, variables included need to be carefully selected; numerical maximization methods need to be carefully designed, and the initial values need to be carefully set. Many authors find it crucial to demean the data before applying the Kalman filter. Mittnik and Zadrozny (2004, p.7) report that "the MLE was not automatic and needed intervention". Aruoba and Scotti (2009) discuss in detail the two steps they use to select the initial values before applying the BFGS numerical maximization. Mariano and Murasawa (2010) use the EM algorithm to obtain an initial estimate and then switch to the quasi-Newton method.

model with mixed or irregular spaced data proposed by Chiu, Eraker, Foerster, Kim, and Seoane (2011, hereafter CEFKS). However, our econometric model and sampling techniques have two genuine differences from those in CEFKS. First, CEFKS assume that lower-frequency data are the result of sampling every k periods from the high frequency variable. In other words, they address a traditional missing data problem without temporal aggregation. Therefore, their posterior conditional distribution of latent states should be an unconstrained multivariate normal distribution. Second, in the CEFKS sampler, single-period (say, date t) latent disaggregated data are drawn conditional on all other latent values. In a VAR(1) model, two neighbors (that is, values in date $t-1$ and $t+1$) are relevant. Though this is a valid sampler, the excess length of the MCMC chain might result in a slow mixing. Our method is to sample latent variables either as a whole or in blocks, where the block size is at the discretion of the user so as to reach a balance between the sampling speed and efficiency.

Lastly, we want to briefly compare our approach with a popular model that handles mixed frequency data, namely the Mi(xed) Da(ta) S(ampling), or MIDAS, regression introduced by Ghysels et al. (2007), Andreou et al. (2010). The MIDAS regression projects high frequency data onto low frequency data with a tightly parameterized weight scheme. Though the MIDAS regression originally focuses on the financial volatility prediction (e.g., Ghysels et al., 2006), it quickly gains popularity among macroeconomists for improving the real-time forecast of key economic variables. See Clements and Galvao (2008, 2009), Marcellino and Schumacher (2010), and Kuzin et al. (2011) for applications.

In the MIDAS regression, the parsimonious declining weights, such as Almon lag polynomial or normalized Beta density, impose a priori structure on the decaying pattern of the regression coefficients. It is true that such a structure prevents parameter proliferation when an aggregation cycle is long, say using daily variables to predict quarterly outcomes. However, for many macroeconomic data the aggregation cycle is relatively short, say monthly-quarterly, or quarterly-annual aggregations. It is both feasible and sensible to adopt a fully data-driven dynamic model like the varied frequency VAR.

In addition, the MIDAS regression is raised mainly in the context of economic forecasting. The VAR, however, can be used for both forecasting and characterizing dynamic relations

among macroeconomic variables. Once the parameters in the reduced-form VAR are routinely estimated, it can be restored to a structural model with suitable identification constraints coming from the economic theory. Furthermore, the VAR with varied frequency data effectively weakens short-run identification assumptions such as zero contemporaneous effects, since it operates on an autoregression with higher frequency data and contemporaneity means a shorter time interval.

The rest of the paper is organized as follows. Section 2.2 specifies of the model. Section 2.3 explains our algorithm to decode latent variables and Section 2.4 compares our method with the Kalman filter solution. Section 2.5 proposes an alternative strategy to sample latent variables block by block. Section 2.6 extends the baseline econometric model to various aggregation types other than simple summation and averaging. Section 2.7 illustrates our approach by a structural VAR model with short-run economic constraints to identify monetary policy shocks. Section 2.8 concludes the paper.

2.2 The model

Assume the k dimensional latent $\{\mathbf{Y}_t^*\}_{t=1}^T$ follow a stationary reduced-form VAR(p) process:

$$\mathbf{Y}_t^* = \mathbf{c} + \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{Y}_{t-i}^* + \varepsilon_t,$$

where $\varepsilon_t \sim N(\mathbf{0}, \mathbf{\Omega})$. The reference time unit is t, which indexes the highest frequency data in the VAR system. The (column) vector $\mathbf{Y}_t^*$ is unobservable since some of the component series may be observed at some lower frequencies and are allowed to change sampling frequencies at any time. Our model is application-oriented and supports varied frequency data of all types. At this level of generality, we must specify a book-keeping convention of observed data. Let $\{Y_t^*\}_{t=1}^T$ be a component series (say, the first variable) in the VAR system. Suppose in some time interval $[a, b]$, $1 \leq a \leq b \leq T$, $a, b \in \mathbb{N}$, disaggregated latent values $Y_a^*, Y_{a+1}^*, ..., Y_{b-1}^*, Y_b^*$ are aggregated into a single observable variate $\overline{Y}_{a,b} \equiv \sum_{j=0}^{b-a} Y_{a+j}^*$. This interval is called an aggregation cycle. We then construct a data series $\{Y_t\}_{t=1}^T$ such that $Y_a = Y_{a+1} = ... = Y_{b-1} = N.A.$ and $Y_b = \overline{Y}_{a,b}$. As a special case, $a = b$ implies the disaggregated value (highest

frequency data) is observed. The data series $\{Y_t\}_{t=1}^T$ contains both observed data and the aggregation structure, since by counting a run of $N.A.$ entries preceding a data point reveals an aggregation cycle. Sometimes high frequency data are grouped into low frequency data by averaging instead of summation, say $\overline{Y}_{a,b} \equiv \frac{1}{b-a+1}\sum_{j=0}^{b-a} Y^*_{a+j}$. In that case, we simply record $Y_b = (b-a+1)\,\overline{Y}_{a,b}$ so that it becomes equivalent to an aggregation by summation. More complicated data averaging types, such as weighed, noisy, missing and nonlinear aggregation, will be discussed in Section 2.6. In fact, this book-keeping convention is a natural way to restore data of different frequencies. Suppose we mostly collect monthly variables but one variable is observed quarterly. In a spreadsheet, it is natural to put these quarterly observations every three entries.

Repeat the above process for each of the k component variables in the VAR system, we obtain k data series. A k-by-T data matrix $\mathbf{Y}$ is constructed by pooling all the data series. Each row is a component data series and each column is the data of k variables at a given time. Clearly $\mathbf{Y}$ contains many $N.A.$ entries, and we define a k-by-T logical matrix $\mathbf{E}$ such that the (i,j) entry of $\mathbf{E}$ equals zero if the corresponding entry of $\mathbf{Y}$ is $N.A.$, and equals one otherwise. The notation $\overrightarrow{\mathbf{E}}$ vectorizes the matrix $\mathbf{E}$ column by column, which will serve as an indexing array to select entries of a matrix (vector) and form a submatrix (subvector). Similarly, the operator $\overrightarrow{\mathbf{Y}}$ vectorizes the matrix $\mathbf{Y}$. Let $\mathbf{Y}^* = (\mathbf{Y}^*_1, ..., \mathbf{Y}^*_T)$, and $\overrightarrow{\mathbf{Y}^*}$ vectorizes the matrix $\mathbf{Y}^*$.

Essentially our tasks are estimating model parameters $\mathbf{\Theta} \equiv \{\mathbf{c}, \mathbf{\Phi}_1, ..., \mathbf{\Phi}_p, \mathbf{\Omega}\}$ and recovering the latent $\mathbf{Y}^*$ from the data matrix $\mathbf{Y}$. A Bayesian framework is adopted since the Gibbs sampler allows separation of these two tasks. For parameter estimation conditional on the complete data, any method handling a standard VAR model applies. For illustration, we treat the VAR as a linear regression by fixing the initial $\mathbf{Y}^*_1, ..., \mathbf{Y}^*_p$. [3]

Denote a 1-by-$(kp+1)$ vector $\mathbf{x}_t = \begin{pmatrix} 1 & \mathbf{Y}^{*\prime}_{t-1} & ... & \mathbf{Y}^{*\prime}_{t-p} \end{pmatrix}$, and a k-by-$(kp+1)$ k block-

[3] Strictly speaking, the VAR model is reduced to a multiple-equation linear regression model only if we neglect the contribution of the initial p observations to the likelihood. Otherwise the posterior conditionals of $\mathbf{c}, \mathbf{\Phi}_1, ..., \mathbf{\Phi}_p$ do not have a closed form. In the classical inference, this is also a popular estimation strategy. Hamilton (1994, p.291) notes that "Vector autoregressions are invariably estimated on the basis of the conditional likelihood function ... rather than the full-sample unconditional likelihood". Maximizing the conditional likelihood is equivalent to OLS regressions equation by equation.

diagonal matrix $\mathbf{X}_t = \begin{pmatrix} \mathbf{x}_t & & \\ & \ddots & \\ & & \mathbf{x}_t \end{pmatrix}$.

Let $\beta = (\mathbf{c}, \mathbf{\Phi}_1, ..., \mathbf{\Phi}_p)'$. With a conjugate proper prior $\overrightarrow{\beta} \sim N(\mu_\beta, \mathbf{V}_\beta)$, $\mathbf{\Omega}^{-1} \sim Wishart(\underline{\mathbf{\Omega}}, \underline{\nu})$, where the operator $\overrightarrow{\beta}$ vectorizes β, we have

$$\overrightarrow{\beta} \mid \cdot \sim N(\mathbf{D}_\beta \mathbf{d}_\beta, \mathbf{D}_\beta),$$
$$\mathbf{\Omega}^{-1} \mid \cdot \sim Wishart\left(\overline{\mathbf{\Omega}}, \overline{\nu}\right),$$

where

$$\mathbf{D}_\beta = \left(\sum_{t=p+1}^{T} \mathbf{X}_t' \mathbf{\Omega}^{-1} \mathbf{X}_t + \mathbf{V}_\beta^{-1} \right)^{-1},$$
$$\mathbf{d}_\beta = \sum_{t=p+1}^{T} \mathbf{X}_t' \mathbf{\Omega}^{-1} \mathbf{Y}_t^* + \mathbf{V}_\beta^{-1} \mu_\beta,$$
$$\overline{\mathbf{\Omega}} = \left[\underline{\mathbf{\Omega}}^{-1} + \sum_{t=p+1}^{T} \left(\mathbf{Y}_t^* - \mathbf{X}_t \overrightarrow{\beta} \right) \left(\mathbf{Y}_t^* - \mathbf{X}_t \overrightarrow{\beta} \right)' \right]^{-1},$$
$$\overline{\nu} = T - p + \underline{\nu}.$$

2.3 Decoding latent variables

In this section, we describe the central step of our sampler, that is, how to sample the latent $\overrightarrow{\mathbf{Y}^*}$ from its posterior conditional distribution. Before presenting our sampler formally, we motivate our approach by a highly simplified scenario of a stationary scalar AR(1):

$$Y_t^* = \phi Y_{t-1}^* + \varepsilon_t, \varepsilon_t \sim N\left(0, \sigma^2\right).$$

The time script t indexes the latent monthly variables. Suppose we only have one quarterly observation $\overline{Y}_{1,3} = Y_1^* + Y_2^* + Y_3^*$ and one monthly observation $Y_4 = Y_4^*$. Conditional on $\phi, \sigma^2, \overline{Y}_{1,3}, Y_4$ we are interested in the posterior distribution of $\overrightarrow{\mathbf{Y}^*} = (Y_1^*, Y_2^*, Y_3^*, Y_4^*)'$. For conciseness, we leave conditioning on ϕ, σ^2 implicit. By our book-keeping convention, $\overrightarrow{\mathbf{Y}} = \left(N.A., N.A., \overline{Y}_{1,3}, Y_4\right)'$, $\overrightarrow{\mathbf{E}} = (0, 0, 1, 1)'$.

First note that suppose $Y_1^* \sim N\left(0, \frac{\sigma^2}{1-\phi^2}\right)$, the stationary distribution of the AR(1) and independent of future disturbances $\varepsilon_2, \varepsilon_3, \varepsilon_4$, then $\overrightarrow{\mathbf{Y}^*} \sim N(\mathbf{0}, \mathbf{\Gamma})$, where the (i,j) entry of $\mathbf{\Gamma}$ equals $\frac{\sigma^2}{1-\phi^2}\phi^{|i-j|}$. However, $\overrightarrow{\mathbf{Y}^*}$ is bound by two linear constraints. First, $Y_1^* + Y_2^* + Y_3^*$ must sum up to the known $\overline{Y}_{1,3}$. Second, Y_4^* must equal to the known Y_4. That implies $\overrightarrow{\mathbf{Y}^*}$ follows a constrained multivariate normal distribution, which can be represented as the product of a conditional normal and a degenerated distribution. To see this, construct a transformation matrix such that

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Then $\mathbf{A}\overrightarrow{\mathbf{Y}^*} = \left(Y_1^*, Y_2^*, \overline{Y}_{1,3}, Y_4\right)' \sim N\left(\mathbf{0}, \widetilde{\mathbf{\Gamma}}\right)$, where $\widetilde{\mathbf{\Gamma}} = \mathbf{A}\mathbf{\Gamma}\mathbf{A}'$.

It follows that

$$(Y_1^*, Y_2^*)' \left| \overline{Y}_{1,3}, Y_4 \sim N\left[\mathbf{\Gamma}_{01}\Gamma_{11}^{-1} \cdot \left(\overline{Y}_{1,3}, Y_4\right)', \mathbf{\Gamma}_{00} - \mathbf{\Gamma}_{01}\Gamma_{11}^{-1}\mathbf{\Gamma}_{10}\right],\right.$$

where $\mathbf{\Gamma}_{01}$ is the submatrix of $\widetilde{\mathbf{\Gamma}}$ with rows selected by $1 - \overrightarrow{\mathbf{E}}$ (i.e., row 1 and 2) and columns selected by $\overrightarrow{\mathbf{E}}$ (i.e., column 3 and 4). The matrixes $\mathbf{\Gamma}_{00}, \mathbf{\Gamma}_{11}, \mathbf{\Gamma}_{10}$ are defined similarly. Practically, in a matrix-based computational environment such as MATLAB, R or GAUSS, a submatrix can be conveniently selected by appending a logical array (indexing vector) to a variable name.

Lastly, the distribution of $\overrightarrow{\mathbf{Y}^*} \left| \overline{Y}_{1,3}, Y_4 \right.$ can be decomposed as the product of $(Y_1^*, Y_2^*)' \left| \overline{Y}_{1,3}, Y_4 \right.$ and $(Y_3^*, Y_4^*)' \left| Y_1^*, Y_2^*, \overline{Y}_{1,3}, Y_4 \right.$, but the latter is degenerated since $Y_3^* = \overline{Y}_{1,3} - Y_1^* - Y_2^*$ and $Y_4^* = Y_4$. So the problem of finding the posterior conditional distribution of $\overrightarrow{\mathbf{Y}^*}$ is resolved.

Now we present the general results on decoding the latent $\overrightarrow{\mathbf{Y}^*}$.

Rewrite the original VAR(p) into a giant VAR(1) system by defining

$$\mathbf{Z}_t^* = \begin{pmatrix} \mathbf{Y}_t^* - \boldsymbol{\mu}_1 \\ ... \\ \mathbf{Y}_{t-p+1}^* - \boldsymbol{\mu}_1 \end{pmatrix}, \mathbf{F} = \begin{pmatrix} \mathbf{\Phi} \\ \mathbf{C} \end{pmatrix}, \mathbf{e}_t = \begin{pmatrix} \varepsilon_t \\ \mathbf{0}_{k(p-1),1} \end{pmatrix}$$

where $\boldsymbol{\mu}_1 = \left(\mathbf{I}_k - \sum_{i=1}^p \mathbf{\Phi}_i\right)^{-1} \mathbf{c}$, $\mathbf{\Phi} = (\mathbf{\Phi}_1, ..., \mathbf{\Phi}_p)$, $\mathbf{C} = \begin{pmatrix} \mathbf{I}_{k(p-1)} & \mathbf{0}_{k(p-1),k} \end{pmatrix}$. So we have

$$\mathbf{Z}_t^* = \mathbf{F}\mathbf{Z}_{t-1}^* + \mathbf{e}_t,$$

and $\mathbf{e}_t \sim N(\mathbf{0}, \boldsymbol{\Delta})$, $\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\Omega} & \\ & \mathbf{0}_{k(p-1),k(p-1)} \end{pmatrix}$, $t = p, ..., T$.

Note that the initial p observations are embodied in the vector $\mathbf{Z}_p^*$.

Proposition 2.1. *Assume the eigenvalues of* $\mathbf{F}$ *all lie inside the unit circle and the initial values* $\mathbf{Z}_p^* \sim N(\mathbf{0}, \mathbf{B})$, *where* $\mathbf{B} = \mathbf{F}\mathbf{B}\mathbf{F}' + \boldsymbol{\Delta}$. *Also assume* $\mathbf{Z}_p^*$ *and future disturbances* $\mathbf{e}_{p+1}, ..., \mathbf{e}_T$ *are independent. Then* $\{\mathbf{Z}_t^*\}_{t=p}^T$ *are strictly stationary with*

$$\begin{pmatrix} \mathbf{Z}_p^* \\ \mathbf{Z}_{p+1}^* \\ \cdots \\ \mathbf{Z}_T^* \end{pmatrix} \sim N\left[\mathbf{0}, \begin{pmatrix} \mathbf{B} & (\mathbf{F}\mathbf{B})' & \cdots & \left(\mathbf{F}^{T-p}\mathbf{B}\right)' \\ \mathbf{F}\mathbf{B} & \mathbf{B} & \cdots & \left(\mathbf{F}^{T-p-1}\mathbf{B}\right)' \\ \vdots & & \ddots & \\ \mathbf{F}^{T-p}\mathbf{B} & \mathbf{F}^{T-p-1}\mathbf{B} & \cdots & \mathbf{B} \end{pmatrix}\right],$$

and $\overrightarrow{\mathbf{Y}^*} \sim N(\boldsymbol{\mu}, \boldsymbol{\Gamma})$, *where* $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', ..., \boldsymbol{\mu}_1')'$. *The* kT*-by-*kT *covariance matrix* $\boldsymbol{\Gamma}$ *is given by*

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_0 & \boldsymbol{\Gamma}_1' & ... & \boldsymbol{\Gamma}_{T-1}' \\ \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_0 & ... & \boldsymbol{\Gamma}_{T-2}' \\ ... & & & \\ \boldsymbol{\Gamma}_{T-1} & \boldsymbol{\Gamma}_{T-2} & ... & \boldsymbol{\Gamma}_0 \end{pmatrix},$$

where $\boldsymbol{\Gamma}_j = E\left[(\mathbf{Y}_t^* - \boldsymbol{\mu}_1)\left(\mathbf{Y}_{t-j}^* - \boldsymbol{\mu}_1\right)'\right] = \sum_{i=1}^p \boldsymbol{\Phi}_i \boldsymbol{\Gamma}_{j-i}$ *with* $\left(\boldsymbol{\Gamma}_{p-1}', ..., \boldsymbol{\Gamma}_0'\right)'$ *being the last* k *columns of* $\mathbf{B}$.

Proof. See appendix. □

On the one hand, the latent $\overrightarrow{\mathbf{Y}^*}$ is regulated by the covariance structure of the VAR(p) process. Proposition 2.1 suggests that the latent $\overrightarrow{\mathbf{Y}^*}$ could be sampled directly from $N(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ if the aggregated data were not known. On the other hand, our knowledge on the aggregated data further sharpens our understandings on $\overrightarrow{\mathbf{Y}^*}$, leading to a constrained multivariate normal distribution which can be decomposed into a conditional normal distribution and a degenerated distribution. The motivating example of AR(1) demonstrates how to connect the aggregated and disaggregated data with a transformation matrix. In the general case, the kT-by-kT transformation matrix $\mathbf{A}$ can be constructed by examining the logical matrix $\mathbf{E}$ that contains

the aggregation structure. First set $\mathbf{A}$ to be a zeros matrix. Then we examine each row of $\mathbf{E}$ and search for the pattern of "a run of zeros ending with a one". Suppose the (i,j) entry of $\mathbf{E}$ is a one preceded by a run of M zeros, we then add $M+1$ ones to $\mathbf{A}$. The locations in $\mathbf{A}$ are row $(j-1)k+i$, column $(j-1)k+i-mk$, $m=0,1,...,M$. Note that $M=0$ is allowed, which simply means a disaggregated value (highest frequency data) is observed.

The new series $\mathbf{A}\overrightarrow{\mathbf{Y}^*}$ transforms the original series $\overrightarrow{\mathbf{Y}^*}$ in such a way that for a $(M+1)$-period temporal aggregation, the first M variates are retained, while the last one is replaced by the sum of the variates in the aggregation cycle. For example, Let $\{Y_t^*\}_{t=1}^T$ be the i^{th} $(i=1,...,k)$ component series in the VAR system. Suppose the aggregation cycle is $[a,a+M]$, $1\leq a\leq T-M$. Then in row $(a-1)k+i,...,(a+M-2)k+i,(a+M-1)k+i$ of $\overrightarrow{\mathbf{Y}^*}$ reside $\left(Y_a^*,...,Y_{a+M-1}^*,Y_{a+M}^*\right)$, while the corresponding entries in $\mathbf{A}\overrightarrow{\mathbf{Y}^*}$ are $\left(Y_a^*,...,Y_{a+M-1}^*,\overline{Y}_{a,a+M}\right)$, where $\overline{Y}_{a,a+M}=\sum_{j=0}^M Y_{a+j}^*$. Conditional on the observed aggregated data $\overline{Y}_{a,a+M}$, the disaggregated variables $\left(Y_a^*,...,Y_{a+M-1}^*\right)$ will follow a conditional normal distribution. To be exact, $\mathbf{A}\overrightarrow{\mathbf{Y}^*}\sim N\left(\mathbf{A}\boldsymbol{\mu},\mathbf{A}\boldsymbol{\Gamma}\mathbf{A}'\right)$, and

$$\overrightarrow{\mathbf{Y}_0^*}\left|\overrightarrow{\mathbf{Y}},\boldsymbol{\Theta}\right.\sim N\left[\boldsymbol{\eta}_0+\boldsymbol{\Gamma}_{01}\boldsymbol{\Gamma}_{11}^{-1}\left(\overrightarrow{\mathbf{Y}_1}-\boldsymbol{\eta}_1\right),\boldsymbol{\Gamma}_{00}-\boldsymbol{\Gamma}_{01}\boldsymbol{\Gamma}_{11}^{-1}\boldsymbol{\Gamma}_{10}\right],$$

where $\overrightarrow{\mathbf{Y}_0^*}$ is the subvector of $\mathbf{A}\overrightarrow{\mathbf{Y}^*}$ selected by the logical vector $1-\overrightarrow{\mathbf{E}}$, namely the latent disaggregated variates. Similarly, $\overrightarrow{\mathbf{Y}_1}$ is the subvector of $\mathbf{A}\overrightarrow{\mathbf{Y}^*}$ selected by $\overrightarrow{\mathbf{E}}$, namely the observed aggregated variates. Where is the realized values of $\overrightarrow{\mathbf{Y}_1}$? They are the subvector of $\overrightarrow{\mathbf{Y}}$ selected by $\overrightarrow{\mathbf{E}}$. As for the conditional mean and variance parameters, $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ are two subvectors of $\mathbf{A}\boldsymbol{\mu}$ selected by $1-\overrightarrow{\mathbf{E}}$ and $\overrightarrow{\mathbf{E}}$ respectively. $\boldsymbol{\Gamma}_{01}$ is the submatrix of $\mathbf{A}\boldsymbol{\Gamma}\mathbf{A}'$ with rows selected by $1-\overrightarrow{\mathbf{E}}$ and columns selected by $\overrightarrow{\mathbf{E}}$. The matrixes $\boldsymbol{\Gamma}_{11},\boldsymbol{\Gamma}_{00},\boldsymbol{\Gamma}_{10}$ are defined similarly.

Note that in the transformation we squeezed out a disaggregated variate at the end of the aggregation cycle, and replaced it with an aggregated data. Those squeezed-out variates, denoted as $\overrightarrow{\mathbf{Y}_{-1}^*}$, correspond to the subvector of $\overrightarrow{\mathbf{Y}^*}$ selected by $\overrightarrow{\mathbf{E}}$. However, $\overrightarrow{\mathbf{Y}_{-1}^*}\left|\overrightarrow{\mathbf{Y}_0^*},\overrightarrow{\mathbf{Y}},\boldsymbol{\Theta}\right.$ is degenerate, since it must equal to the difference between the aggregated value and the sum of the rest disaggregated values.

The conditional normal $\overrightarrow{\mathbf{Y}_0^*}\left|\overrightarrow{\mathbf{Y}},\boldsymbol{\Theta}\right.$ plus the degenerated $\overrightarrow{\mathbf{Y}_{-1}^*}\left|\overrightarrow{\mathbf{Y}_0^*},\overrightarrow{\mathbf{Y}},\boldsymbol{\Theta}\right.$ fully characterize

the distribution of $\overrightarrow{\mathbf{Y}^*}\left|\overrightarrow{\mathbf{Y}},\mathbf{\Theta}\right.$, for $\left(\overrightarrow{\mathbf{Y}_0^*},\overrightarrow{\mathbf{Y}^*}\right)$ is a partition of $\overrightarrow{\mathbf{Y}^*}$ (note that $\overrightarrow{\mathbf{Y}_0^*}$ has another identity, namely the subvector of $\overrightarrow{\mathbf{Y}^*}$ selected by $1-\overrightarrow{\mathbf{E}}$). That finishes the Gibbs sampler to the latent variables.

As a special case, the above procedure accommodates one-period aggregation, in which a disaggregated value (highest frequency data) is observed. Let Y_t^* be such an observed disaggregated value. In the step of calculating $\overrightarrow{\mathbf{Y}_0^*}\left|\overrightarrow{\mathbf{Y}},\mathbf{\Theta}\right.$, the variable Y_t^* is squeezed out from $\overrightarrow{\mathbf{Y}_0^*}$, but counted as a member of $\overrightarrow{\mathbf{Y}}$ to sharpen our understandings on other latent variables. In the step of calculating $\overrightarrow{\mathbf{Y}_{-1}^*}\left|\overrightarrow{\mathbf{Y}_0^*},\overrightarrow{\mathbf{Y}},\mathbf{\Theta}\right.$, the variable Y_t^* belongs to $\overrightarrow{\mathbf{Y}_{-1}^*}$ and equals to the realized value of itself.

2.4 A comparison with the Kalman filter

The state space framework with the Kalman filter is powerful enough to bridge any frequency mismatch. Kalman filter can handle temporal aggregation for the sure, though it is not necessarily the best solution to the current problem.

Consider again the AR(1) example with one quarterly observation $\overline{Y}_{1,3}$ and one monthly observation Y_4 as in Section 2.3. The state space representation of that model consists of a transition equation and a measurement equation such that

$$\boldsymbol{\xi}_t = \mathbf{G}\boldsymbol{\xi}_{t-1} + \mathbf{u}_t,$$

$$z_t = \mathbf{H}_t\boldsymbol{\xi}_t.$$

where in the transition equation the states $\boldsymbol{\xi}_t = \left(Y_t^*, Y_{t-1}^*, Y_{t-2}^*\right)'$, $\mathbf{G} = \begin{pmatrix} \phi & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, $\mathbf{u}_t = (\varepsilon_t, 0, 0)'$, $t = 1,2,3,4$; In the measurement equation, $z_3 = \overline{Y}_{1,3}$, $z_4 = Y_4$, $\mathbf{H}_3 = (1,1,1)$, $\mathbf{H}_4 = (1,0,0)$ while $z_1, z_2, \mathbf{H}_1, \mathbf{H}_2$ are empty[4]. The initial distribution of $\boldsymbol{\xi}_0$ is conformable with the stationary AR(1) process, that is, $\boldsymbol{\xi}_0 \sim N(\mathbf{0}, \mathbf{B})$, where $\mathbf{B} = \mathbf{G}\mathbf{B}\mathbf{G}' + \mathbf{\Delta}$, and $\mathbf{\Delta} = diag\left(\sigma^2, 0, 0\right)$.

[4]Alternatively, empty state vector can be circumvented by letting z_1, z_2 be some exogenous random variables whose data generating processes are unrelated with model parameters so that the likelihood is only shifted by a constant (see Mariano and Murasawa, 2003). The only advantage of introducing such artificial random variables is to keep constant the size of the measurement vector.

One way to describe the Kalman filter is to break the recursion into predicting and updating steps. Due to no information in date 1 and 2, the updating step are skipped and both $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ follow $N(\mathbf{0}, \mathbf{B})$. Then predict $\boldsymbol{\xi}_3 \sim N(\mathbf{0}, \mathbf{B})$, $z_3 \sim N(\mathbf{0}, \mathbf{H_3 B H_3'})$ and use realized z_3 to update $\boldsymbol{\xi}_3 | z_3 \sim N\left(\widehat{\boldsymbol{\xi}}_{3|3}, \mathbf{P}_{3|3}\right)$, where

$$\begin{aligned}
\widehat{\boldsymbol{\xi}}_{3|3} &= \mathbf{B}\mathbf{H}_3' \left(\mathbf{H_3 B H_3'}\right)^{-1} z_3, \\
\mathbf{P}_{3|3} &= \mathbf{B} - \mathbf{B}\mathbf{H}_3' \left(\mathbf{H_3 B H_3'}\right)^{-1} \mathbf{H_3 B}.
\end{aligned}$$

In date 4, first predict $\boldsymbol{\xi}_4 | z_3 \sim N\left(\widehat{\boldsymbol{\xi}}_{4|3}, \mathbf{P}_{4|3}\right)$ and $z_4 | z_3 \sim N\left(\mathbf{H}_4 \widehat{\boldsymbol{\xi}}_{4|3}, \mathbf{H}_4 \mathbf{P}_{4|3} \mathbf{H}_4'\right)$, where

$$\begin{aligned}
\widehat{\boldsymbol{\xi}}_{4|3} &= \mathbf{G}\widehat{\boldsymbol{\xi}}_{3|3}, \\
\mathbf{P}_{4|3} &= \mathbf{G}\mathbf{P}_{3|3}\mathbf{G}' + \boldsymbol{\Delta}.
\end{aligned}$$

Then use realized z_4 to update $\boldsymbol{\xi}_4 | z_3, z_4 \sim N\left(\widehat{\boldsymbol{\xi}}_{4|4}, \mathbf{P}_{4|4}\right)$, where

$$\begin{aligned}
\widehat{\boldsymbol{\xi}}_{4|4} &= \widehat{\boldsymbol{\xi}}_{4|3} + \mathbf{P}_{4|3}\mathbf{H}_4' \left(\mathbf{H}_4 \mathbf{P}_{4|3} \mathbf{H}_4'\right)^{-1} \left(z_4 - \mathbf{H}_4 \widehat{\boldsymbol{\xi}}_{4|3}\right), \\
\mathbf{P}_{4|4} &= \mathbf{P}_{4|3} - \mathbf{P}_{4|3}\mathbf{H}_4' \left(\mathbf{H}_4 \mathbf{P}_{4|3} \mathbf{H}_4'\right)^{-1} \mathbf{H}_4 \mathbf{P}_{4|3}.
\end{aligned}$$

After the long recursion, we obtain the likelihood function, which is the product of the densities of z_3 and $z_4 | z_3$. Note that the original system is a scalar AR(1), but in the state space form we expand the state vector $\boldsymbol{\xi}_t$ to three dimensions with a 3-by-3 transition matrix the $\mathbf{G}$, which consists of only one "material element" ϕ but many "auxiliary elements" like zeros and ones. Furthermore, the recursive formula yields an illusion that the likelihood function is in a complicated form. However, the explicit likelihood is not only tractable but also simple. The spirit of our approach is to explore the joint distribution of latent variables of multiple periods. Since $(Y_1^*, Y_2^*, Y_3^*, Y_4^*)'$ is multivariate normal, by a linear transformation $\left(Y_1^*, Y_2^*, \overline{Y}_{1,3}, Y_4\right)'$ is also multivariate normal with $N\left(\mathbf{0}, \widetilde{\boldsymbol{\Gamma}}\right)$, where $\widetilde{\boldsymbol{\Gamma}}$ has been defined in Section 2.3. The likelihood function of the observed variates $\left(\overline{Y}_{1,3}, Y_4\right)'$ is a multivariate normal density with the covariance matrix given by the last two rows and columns of $\widetilde{\boldsymbol{\Gamma}}$.

Similarly, to find the posterior distribution of the latent states, the Kalman filter offers a smoothing recipe. $\boldsymbol{\xi}_4 | z_3, z_4$ has already been obtained in the forward recursion. $\boldsymbol{\xi}_3 | \boldsymbol{\xi}_4, z_3, z_4$

has the same distribution as $\boldsymbol{\xi}_3 | \boldsymbol{\xi}_4, z_3$, that is $N\left(\widehat{\boldsymbol{\xi}}_{3|4}, \mathbf{P}_{3|4}\right)$, where

$$\widehat{\boldsymbol{\xi}}_{3|4} = \widehat{\boldsymbol{\xi}}_{3|3} + \mathbf{P}_{3|3}\mathbf{G}'\left(\mathbf{P}_{4|3}\right)^{-1}\left(\boldsymbol{\xi}_4 - \widehat{\boldsymbol{\xi}}_{4|3}\right),$$
$$\mathbf{P}_{3|4} = \mathbf{P}_{3|3} - \mathbf{P}_{3|3}\mathbf{G}'\left(\mathbf{P}_{4|3}\right)^{-1}\mathbf{G}\mathbf{P}_{3|3}.$$

The distribution of $\boldsymbol{\xi}_4 | z_3, z_4$ and $\boldsymbol{\xi}_3 | \boldsymbol{\xi}_4, z_3, z_4$ are enough to characterize the posterior distribution of $(Y_1^*, Y_2^*, Y_3^*, Y_4^*)'$. However, a technical problem is that the covariance matrix of $\boldsymbol{\xi}_4 | z_3, z_4$ is not of full rank (the first column is zeros) and the covariance matrix of $\boldsymbol{\xi}_3 | \boldsymbol{\xi}_4, z_3, z_4$ is an entire zeros matrix. This is caused by two facts: i) some components in the state vector are actually observed; ii) state vectors of different periods have overlapping components. The problem can be solved by deleting the non-random components, though it inevitably increases the implementation complexity. The advantage of our approach is the transparency of the posterior distribution. As is stated explicitly in Section 2.3, the posterior $(Y_1^*, Y_2^*, Y_3^*, Y_4^*)'$ follows a multivariate normal distribution subject to two constraints, which can be decomposed into a bivariate normal in terms of $(Y_1^*, Y_2^*)' \left| \overline{Y}_{1,3}, Y_4 \right.$ and the degenerated $(Y_3^*, Y_4^*)' \left| Y_1^*, Y_2^*, \overline{Y}_{1,3}, Y_4 \right.$.

In a general VAR(p) model with linear temporal aggregation of any types, the explicit form of the likelihood still exists. The disaggregated variables follow a joint multivariate normal distribution conformable to the VAR process. The observed variables are linear combinations of those disaggregated variables. Since the normality is preserved under linear transformations, the likelihood function of observed variates will take the form of a multivariate normal density. Actually its form has already given in Section 2.3. Note that $\mathbf{A}\overrightarrow{\mathbf{Y}^*} \sim N\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}'\right)$ and the observed variates are those elements of $\mathbf{A}\overrightarrow{\mathbf{Y}^*}$ selected by $\overrightarrow{\mathbf{E}}$. It follows that the likelihood function is a multivariate normal density with mean $\boldsymbol{\eta}_1$ and covariance matrix $\boldsymbol{\Gamma}_{11}$.

2.5 Gibbs sampler with Blocks

The sampling procedure in Section 2.3 allows us to draw the latent variables all at once. However, if we have a large dataset, the procedure requires manipulating large matrixes and their inversions, which poses computational challenges. In this section we propose an alternative sampler which divides latent variables into blocks. It speeds up the sampler and saves computer memories, at the price of increasing the length of the MCMC chain. The idea of this approach

is to explore the Markov property to simplify conditional distributions and thus reduce the size of matrixes. The following proposition explains the (high order) Markov property of the VAR(p) process.

Proposition 2.2. *Suppose* $\{\mathbf{Y}_t^*\}_{t=1}^T$ *follow a VAR(p) process, then*

$$\begin{aligned} & p\left(\mathbf{Y}_s^*, ..., \mathbf{Y}_t^* \left| \mathbf{Y}_1^*, ..., \mathbf{Y}_{s-1}^*, \mathbf{Y}_{t+1}^*, ..., \mathbf{Y}_T^* \right.\right) \\ & = p\left(\mathbf{Y}_s^*, ..., \mathbf{Y}_t^* \left| \mathbf{Y}_{s-p}^*, ..., \mathbf{Y}_{s-1}^*, \mathbf{Y}_{t+1}^*, ..., \mathbf{Y}_{t+p}^* \right.\right), \end{aligned}$$

for all $1 \leq s \leq t \leq T$ *and* $s, t \in \mathbb{N}$. *Over-ranged variables* $\mathbf{Y}_0^*, \mathbf{Y}_{-1}^*, ...$ *and* $\mathbf{Y}_{T+1}^*, \mathbf{Y}_{T+2}^*, ...$ *are treated as N.A..*

Proof. See appendix. □

To fix ideas on how to use Proposition 2.2 in the Gibbs sampler with blocks, consider again the AR(1) example in Section 2.3:

$$Y_t^* = \phi Y_{t-1}^* + \varepsilon_t, \varepsilon_t \sim N\left(0, \sigma^2\right).$$

Suppose we observe two quarterly observations $\overline{Y}_{1,3} = Y_1^* + Y_2^* + Y_3^*$, $\overline{Y}_{4,6} = Y_4^* + Y_5^* + Y_6^*$ and one monthly observation $Y_7 = Y_7^*$. Conditional on $\phi, \sigma^2, \overline{Y}_{1,3}, \overline{Y}_{4,6}, Y_7$ we are interested in the posterior distribution of $\overrightarrow{\mathbf{Y}^*} = (Y_1^*, ..., Y_7^*)'$. If we sample the latent variates all at once, we need to work on a 7-by-7 transformation matrix and covariance matrix.

Now we partition the seven latent variables into three blocks and consecutively sample variates in one block conditional on other two blocks. We need to specify the posterior conditionals of

$$\begin{aligned} & (Y_1^*, Y_2^*, Y_3^*)' \left| \overline{Y}_{1,3}, \overline{Y}_{4,6}, Y_7, Y_4^*, Y_5^*, Y_6^*, Y_7^* \right., \\ & (Y_4^*, Y_5^*, Y_6^*)' \left| \overline{Y}_{1,3}, \overline{Y}_{4,6}, Y_7, Y_1^*, Y_2^*, Y_3^*, Y_7^* \right., \\ & Y_7^* \left| \overline{Y}_{1,3}, \overline{Y}_{4,6}, Y_7, Y_1^*, ..., Y_6^* \right.. \end{aligned}$$

First, to sample $(Y_1^*, Y_2^*, Y_3^*)'$ conditional on all other variables, we note two facts: i) conditioning on $\overline{Y}_{4,6}$ and Y_7 is redundant since we already know $Y_4^*, Y_5^*, Y_6^*, Y_7^*$; ii) the Markov property of the AR(1) process implies that once we know Y_4^*, further knowledge on Y_5^*, Y_6^*, Y_7^*

is irrelevant. So it is equivalent to work on $(Y_1^*, Y_2^*, Y_3^*)' \left| \overline{Y}_{1,3}, Y_4^* \right.$. By a linear transformation, $\left(Y_1^*, Y_2^*, \overline{Y}_{1,3}, Y_4^*\right)'$ follows a multivariate normal, so we first sample $(Y_1^*, Y_2^*)'$ conditional on $\overline{Y}_{1,3}, Y_4^*$, and then sample the degenerated Y_3^*, which is essentially the same as the example in Section 2.3. This process only requires a 4-by-4 transformation matrix and covariance matrix.

Next, we consider sampling $(Y_4^*, Y_5^*, Y_6^*)'$ conditional on all other variables. Again we note two facts: i) Our knowledge on Y_1^*, Y_2^*, Y_3^* makes $\overline{Y}_{1,3}$ redundant; ii) the Markov property implies that conditioning on Y_3^*, Y_7^* is sufficient. So we work on $(Y_4^*, Y_5^*, Y_6^*)' \left| \overline{Y}_{4,6}, Y_3^*, Y_7 \right.$ by first sampling $(Y_4^*, Y_5^*)' \left| \overline{Y}_{4,6}, Y_3^*, Y_7 \right.$ from a conditional normal distribution and then sampling the degenerated Y_6^*. This process only requires a 5-by-5 transformation matrix and covariance matrix.

Third, $Y_7^* \left| \overline{Y}_{1,3}, \overline{Y}_{4,6}, Y_7, Y_1^*, ..., Y_6^* \right.$ is degenerated since we know its realization.

This example illustrates how we simplify the posterior conditional distribution by the Markov property, though the computational advantage is mild in this case. Now we extend this example by assuming 100 quarterly observations, say $\overline{Y}_{3j-2,3j}$, $j = 1, ..., 100$. Treat variables in each aggregation cycle as a block, that is, $\left(Y_{3j-2}^*, Y_{3j-1}^*, Y_{3j}^*\right)$, $j = 1, ..., 100$. To sample each block conditioning on all other variables, we work on $Y_{3j-2}^*, Y_{3j-1}^*, Y_{3j}^* \left| \overline{Y}_{3j-2,3j}, Y_{3j-3}^*, Y_{3j+1}^* \right.$ (with proper adjustment for first and last blocks). First sample $Y_{3j-2}^*, Y_{3j-1}^*, \overline{Y}_{3j-2,3j} \left| Y_{3j-3}^*, Y_{3j+1}^* \right.$ from a conditional normal distribution and then recover Y_{3j}^* by a direct subtraction. Proposition 2.1 suggests the joint distribution of $\left(Y_{3j-3}^*, Y_{3j-2}^*, Y_{3j-1}^*, Y_{3j}^*, Y_{3j+1}^*\right)$ is identical for all $j = 2, ..., 99$ due to strict stationarity, which also implies the same form of the conditional normal distribution can be applied to all the 98 variable blocks. This feature substantially accelerates the sampler since most computation time is spend on calculating the conditional variances.

Also note that the size of the block is not necessarily set to an aggregation cycle. For example, it is legitimate to partition the series every two aggregation cycles. If the entire series are treated as one block, we go back to the sampler specified in Section 2.3. Generally speaking, larger block size increases computation time, but also reduces the length and improves the mixing of the MCMC chain. The balance between the sampling efficiency and speed is at the discretion of practitioners.

In a general VAR(p)-based model, the blocking strategy still applies. Though the variable blocks can be set arbitrarily, it is natural to set the dividing lines at the end of the aggregation cycle. Let $\mathbf{E}, \mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\Gamma}$ as defined in Section 2.3. Suppose we intend to sample $\left(\mathbf{Y}_t^*, \mathbf{Y}_{t+2}^*..., \mathbf{Y}_{t+j}^*\right)$ in one block, conditional on disaggregated draws of all other blocks. The Markov property of the VAR(p) model suggests that we only need to use the joint distribution of $\left(\mathbf{Y}_{t-p}^*, ..., \mathbf{Y}_t^*..., \mathbf{Y}_{t+j}^*, ..., \mathbf{Y}_{t+j+p}^*\right)$ to formulate the conditional normal distribution. Let $\mathbf{E}_0$ be a k-by-T logical matrix of zeros, except that column t to $t+j$ are equal to the corresponding columns in $1-\mathbf{E}$. Let $\mathbf{E}_1$ be a k-by-T logical matrix of zeros, except that column $t-p$ to column $t-1$ are ones, column t to $t+j$ are identical to the corresponding columns in $\mathbf{E}$, column $t+j+1$ to column $t+j+p$ are ones. Then we can use the vectorized $\overrightarrow{\mathbf{E}}_0$ and $\overrightarrow{\mathbf{E}}_1$ to select submatrixes to form the conditional normal distribution:

$$\overrightarrow{\mathbf{Y}_0^*} \left| \overrightarrow{\mathbf{Y}_1^*}, \overrightarrow{\mathbf{Y}}, \boldsymbol{\Theta} \sim N\left[\boldsymbol{\eta}_0 + \boldsymbol{\Gamma}_{01}\boldsymbol{\Gamma}_{11}^{-1}\left(\overrightarrow{\mathbf{Y}_1^*} - \boldsymbol{\eta}_1\right), \boldsymbol{\Gamma}_{00} - \boldsymbol{\Gamma}_{01}\boldsymbol{\Gamma}_{11}^{-1}\boldsymbol{\Gamma}_{10}\right],\right.$$

where $\overrightarrow{\mathbf{Y}_0^*}, \overrightarrow{\mathbf{Y}_1^*}$ are the subvectors of $\mathbf{A}\overrightarrow{\mathbf{Y}^*}$ selected by the logical vectors $\overrightarrow{\mathbf{E}}_0$ and $\overrightarrow{\mathbf{E}}_1$ respectively, $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ are two subvectors of $\mathbf{A}\boldsymbol{\mu}$ selected by $\overrightarrow{\mathbf{E}}_0$ and $\overrightarrow{\mathbf{E}}_1$ respectively. $\boldsymbol{\Gamma}_{01}$ is the submatrix of $\mathbf{A}\boldsymbol{\Gamma}\mathbf{A}'$ with rows selected by $\overrightarrow{\mathbf{E}}_0$ and columns selected by $\overrightarrow{\mathbf{E}}_1$. The matrixes $\boldsymbol{\Gamma}_{11}, \boldsymbol{\Gamma}_{00}, \boldsymbol{\Gamma}_{10}$ are defined similarly. Lastly, the squeezed out disaggregated values can be recovered by a direct subtraction.

2.6 Other aggregation types

Macroeconomic data may exhibit more complicated aggregation types other than summation and simple average. In this section, we extend the model to various aggregation types that an empirical researcher may encounter.

2.6.1 Weighed aggregation

In the baseline model, the time interval $[t, t+1]$ is equidistant over time. However, calender days vary in a month, and working days are affected by holidays. Suppose latent daily values of some variable are simple-averaged to generate latent quarterly data and observable annual data. Assume there are $66, 66, 66, 60$ working days in the four quarters, and then the latent quarterly

values $\{Y_t^*\}_{t=1}^T$ are linked to the annual data $\left\{\overline{Y}_{4i+1,4i+4}\right\}_{i=0}^{T/4-1}$ by the relation $\overline{Y}_{4i+1,4i+4} = \frac{66}{252}Y_{4i+1}^* + \frac{66}{252}Y_{4i+2}^* + \frac{66}{252}Y_{4i+3}^* + \frac{60}{252}Y_{4i+4}^*$.

In a general setting, let $\{Y_t^*\}_{t=1}^T$ be a component series in the VAR system. Suppose in an aggregation cycle $[a,b]$, disaggregated latent values $Y_a^*, Y_{a+1}^* ... Y_{b-1}^*, Y_b^*$ are aggregated into an observed data $\overline{Y}_{a,b} \equiv \sum_{j=0}^{b-a} \omega_{a+j} Y_{a+j}^*$, where $\{\omega_t\}_{t=1}^T$ is a deterministic weight series. In the above example, the weight series looks like $\left\{..., \frac{66}{252}, \frac{66}{252}, \frac{66}{252}, \frac{60}{252}, ...\right\}$. The data series $\{Y_t\}_{t=1}^T$ and the matrixes $\mathbf{Y}, \mathbf{Y}^*, \mathbf{E}$ are constructed in the same way as in Section 2.3, but the transformation matrix $\mathbf{A}$ needs to incorporate the weights information. The matrix $\mathbf{A}$ can be constructed based on a kT-by-kT matrix of zeros. Then we examine the logical matrix $\mathbf{E}$ row by row to modify $\mathbf{A}$ as appropriate. Suppose the (i,j) entry of $\mathbf{E}$ is a one preceded by a run of M zeros. By reading the weight series $\{\omega_t\}_{t=1}^T$ of variable i, we extract $\omega_j, \omega_{j-1}, ..., \omega_{j-M}$ and place them to $\mathbf{A}$. The locations in $\mathbf{A}$ are row $(j-1)\,k+i$, column $(j-1)\,k+i-mk$, $m = 0, ..., M$. The rest sampling procedure remains the same.

2.6.2 Differenced data and weighed aggregation

The VAR model in use is stationary. However, many macroeconomic variables contain unit roots. It is common to put the first-differenced variables in the VAR system, though in the current model cointegration relations and error correction terms are not included. The model with cointegration is left for future research.

Consider an example. Let $\{Y_t^*\}_{t=1}^T$ be the latent monthly GDP series and we actually put $\Delta Y_t^* \equiv Y_t^* - Y_{t-1}^*$ as a component variable in the VAR model. We observe the quarterly GDP series $\overline{Y}_{t,t+2} = Y_t^* + Y_{t+1}^* + Y_{t+2}^*$, $t = 1, 4, 7, ...$ Define the quarterly-differenced data $\Delta^3\overline{Y}_{t,t+2} = \overline{Y}_{t,t+2} - \overline{Y}_{t-3,t-1}$. The observable quarterly-differenced data and the unobservable monthly-differenced data are linked by the relation

$$\begin{aligned}
\Delta^3\overline{Y}_{t,t+2} &= \sum_{j=0}^{2} \left(Y_{t+j}^* - Y_{t-3+j}^*\right) \\
&= \sum_{j=0}^{2} \left(\Delta Y_{t+j}^* + \Delta Y_{t-1+j}^* + \Delta Y_{t-2+j}^*\right) \\
&= \Delta Y_{t+2}^* + 2\Delta Y_{t+1}^* + 3\Delta Y_t^* + 2\Delta Y_{t-1}^* + \Delta Y_{t-2}^*.
\end{aligned}$$

In other words, the observed quarterly GDP growth series is a weighed sum of the unobserved monthly GDP growth series. Similarly, suppose the aggregated value is formed by taking average instead of summation, that is, $\overline{Y}_{t,t+2} = \frac{1}{3}\left(Y_t^* + Y_{t+1}^* + Y_{t+2}^*\right)$. The quarterly-monthly differenced data are linked by the relation

$$\Delta^3\overline{Y}_{t,t+2} = \frac{1}{3}\Delta Y_{t+2}^* + \frac{2}{3}\Delta Y_{t+1}^* + \Delta Y_t^* + \frac{2}{3}\Delta Y_{t-1}^* + \frac{1}{3}\Delta Y_{t-2}^*.$$

In principle, the approach handling weighed averaged data in the previous subsection applies to the current problem. We just put quarterly differenced data in the data series $\{Y_t\}_{t=1}^T$ every three entries (with other entries being $N.A.$). As for the transformation matrix $\mathbf{A}$, suppose we are reading row i and column j of $\mathbf{E}$ and (i,j) is a non-zero entry. Then for summation-type aggregation we place $(1,2,3,2,1)$ to $\mathbf{A}$ at row $(j-1)\,k+i$ and column $(j-1)\,k+i-mk$, $m=0,...,4$. The rest sampling procedure remains the same.

In practice, it is preferable to estimate the differenced-data model using a block Gibbs sampler.[5] However, in this case the block sampler differs slightly from the previous one, since the aggregation of differenced data spans across two aggregation cycles. This feature implies that two aggregated data are relevant when we sample a block of variables in an aggregation cycle.

For illustration, consider again a monthly AR(1) model such that

$$\Delta Y_t^* = \phi \cdot \Delta Y_{t-1}^* + \varepsilon_t,$$

while only quarterly-differenced variables $\Delta^3\overline{Y}_{t,t+2},\ t = 4,7,10,...$ are observed. Suppose we intend to sample the block $\left(\Delta Y_t^*, \Delta Y_{t+1}^*, \Delta Y_{t+2}^*\right)$ conditional on all the other monthly-differenced data as well as quarterly-differenced data. In this case, two aggregated values

[5] It seems that there are some numerical issues if the disaggregated, differenced data are sampled all at once. Consider a univariate AR(1) with $\phi = 0.5, \sigma^2 = 1$. Let the covariance matrix of T observations be $\mathbf{\Gamma}$, where the (i,j) entry of $\mathbf{\Gamma}$ equals $\frac{\sigma^2}{1-\phi^2}\phi^{|i-j|}$. Construct the transformation matrix $\mathbf{A}$ with weights $(1,2,3,2,1)$ assigned as appropriate. The transformed covariance matrix $\mathbf{A\Gamma A}'$ is positive definite in theory. However, it seems that when T is larger than 100, MATLAB cannot perform the cholesky decomposition and produces a non-positive definite error, though we know theoretically the cholesky factor in this case is $\mathbf{AL}$, where $\mathbf{LL}' = \mathbf{\Gamma}$. The puzzle is that regardless of T MATLAB can always cholesky decompose $\mathbf{\Gamma}$ and $\mathbf{A\Gamma A}'$ for level-data aggregation specified in the previous section. We are not aware of the source of this numerical problem, so currently we estimate the differenced-data model using the blocking strategy, in which the transformed covariance matrix is relatively small in size and no obvious numerical problem is detected.

$\Delta^3\overline{Y}_{t,t+2}$, $\Delta^3\overline{Y}_{t+3,t+5}$ bind disaggregated $\left(\Delta Y_t^*, \Delta Y_{t+1}^*, \Delta Y_{t+2}^*\right)$ such that

$$\Delta Y_{t+2}^* + 2\Delta Y_{t+1}^* + 3\Delta Y_t^* = \Delta^3\overline{Y}_{t,t+2} - 2\Delta Y_{t-1}^* - \Delta Y_{t-2}^*,$$
$$2\Delta Y_{t+2}^* + \Delta Y_{t+1}^* = \Delta^3\overline{Y}_{t+3,t+5} - \Delta Y_{t+5}^* - 2\Delta Y_{t+4}^* - 3\Delta Y_{t+3}^*.$$

In others words, $\left(\Delta Y_t^*, \Delta Y_{t+1}^*, \Delta Y_{t+2}^*\right)$ follows a condition normal distribution subject to two linear constraints. So we first explore the Markov property of the AR(1) process and use the (unconditional) joint normal distribution of $\left(\Delta Y_{t-1}^*, \Delta Y_t^*, \Delta Y_{t+1}^*, \Delta Y_{t+2}^*, \Delta Y_{t+3}^*\right)$ to find out the distribution of $\left(\Delta Y_t^*, \Delta Y_{t+1}^*, \Delta Y_{t+2}^*\right)$ conditional on all the other disaggregated data. Then we build a transformation matrix with the purpose of taking $\left(\Delta Y_t^*, \Delta Y_{t+1}^*, \Delta Y_{t+2}^*\right)$ into $\left(\Delta Y_{t+2}^* + 2\Delta Y_{t+1}^* + 3\Delta Y_t^*, \Delta Y_{t+1}^*, 2\Delta Y_{t+2}^* + \Delta Y_{t+1}^*\right)$. Note that the first term has a realized value $\Delta^3\overline{Y}_{t,t+2} - 2\Delta Y_{t-1}^* - \Delta Y_{t-2}^*$, and the third term has its realization $\Delta^3\overline{Y}_{t+3,t+5} - \Delta Y_{t+5}^* - 2\Delta Y_{t+4}^* - 3\Delta Y_{t+3}^*$, so we first sample $\Delta Y_{t+1}^* \left| \left(\Delta Y_{t+2}^* + 2\Delta Y_{t+1}^* + 3\Delta Y_t^*\right), \left(2\Delta Y_{t+2}^* + \Delta Y_{t+1}^*\right)\right.$ and then sample the degenerated ΔY_{t+2}^*, $3\Delta Y_t^*$.

In a general VAR(p)-based model with differenced data, the blocking strategies still applies. First, choose a block size. Second, use Markov property of the VAR(p) to find out the distribution of disaggregated variables within the block conditional on all the other disaggregated variables. Third, find out all the aggregation constraints that bind the variables within the block. Fourth, make linear transformations to accommodate those constraints. Fifth, sample disaggregated variables from a conditional normal distribution and degenerated distribution.

2.6.3 Data revision and noisy aggregation

If the VAR model is mainly used for real-time forecasting, it is necessary to incorporate all the recent data. However, some latest macroeconomic data might be less accurate and subject to revision. In that case the most recent aggregated data might be viewed as the summation of the latent disaggregated values plus a noise. The noisy aggregation can be modeled as follows. Let $\{Y_t^*\}_{t=1}^T$ be a component series. Suppose in some time interval $[a, b]$ disaggregated latent values $Y_a^*, Y_{a+1}^* ... Y_{b-1}^*, Y_b^*$ are grouped into an aggregated observed data $\overline{Y}_{a,b} \equiv u_a + \sum_{j=0}^{b-a} Y_{a+j}^*$, where u_a follows an independent $N(0, \eta)$ regardless of the time script a. At a later stage, the authority revises the aggregated data so as to remove the noise u_a. In other

words, historical noises are known, leaving only the latest noise unknown. Suppose a researcher has realizations of historical noises $\{u_a\}_{a=1}^{J}$ in hand. With a conjugate prior $\eta \sim IG(c_1, c_2)$, the posterior conditional distribution is $\eta \sim IG\left[\frac{J}{2} + c_1, \left(c_2^{-1} + \frac{1}{2}\sum_{a=1}^{J} u_a^2\right)^{-1}\right]$. To sample the latent disaggregated values from their posterior conditional distribution, we take the previous draw of η as given. The data series $\{Y_t\}_{t=1}^{T}$ are constructed by filling in the corresponding entries with revised data except for the most recent one with noise-ridden data. Transformation matrix is constructed as usual, but we modify the covariance matrix of $\mathbf{A}\overrightarrow{\mathbf{Y}^*}$, which is $\mathbf{A\Gamma A}'$ originally. Suppose this noise-ridden data happens to variable i at date j. By adding the $((j-1)k+i, (j-1)k+i)$ entry of $\mathbf{A\Gamma A}'$ by η, we obtain the new covariance matrix of $\mathbf{A}\overrightarrow{\mathbf{Y}^*}$. The rest sampling procedure remains the same.

2.6.4 Missing data and no aggregation

Though the missing data problem is more common in survey, industrial or regional data at micro level, missing macroeconomic data may be present in the oldest or latest data. Consider the real-time forecasting again. Many economic indicators are published with a time lag. At the time when a forecasting must be made, some latest variables may be available while some are not, hence the missing data. Our model can conveniently handle missing data by classifying them into latent disaggregated variates block. In the data matrix $\mathbf{Y}$, record the missing data as, say, $M.S.$. Then define logical matrix $\mathbf{E}$ such that the (i, j) entry in $\mathbf{E}$ equals zero if the corresponding entry in $\mathbf{Y}$ is $N.A.$, and equals one if that in $\mathbf{Y}$ contains data, and equals two if that in $\mathbf{Y}$ is $M.S.$. The construction of the transformation matrix $\mathbf{A}$ still starts from an identity matrix and we modify it by examining $\mathbf{E}$. If the (i, j) entry of $\mathbf{E}$ entry is zero or two, skip and proceed to column $j+1$ (or conclude this row). Otherwise, we search column $j-1, j-2, ...$ for a run of zeros and insert ones into $\mathbf{A}$ in the same way as before. Once the transformation matrix is constructed, replace all the twos with zeros in $\mathbf{E}$ and use it to select a submatrix or subvector. The rest sampling procedure remains the same.

2.6.5 Logarithmic data and nonlinear aggregation

Using logarithmic variables in the VAR has many merits, but it also introduces nonlinearity in the aggregation structure. Our model is a linear model that handles temporal aggregation by exploring the fact that normality is preserved under linear transformations. Suppose three monthly data are averaged into a quarterly data such that $\overline{Y}_{1,3} = \frac{1}{3}(Y_1^* + Y_2^* + Y_3^*)$. If logarithmic monthly variable $(\ln Y_1^*, \ln Y_2^*, \ln Y_3^*)$ are used in the VAR system, they follow multivariate normal but conditional on $\overline{Y}_{1,3}$ they do not, for $\ln \overline{Y}_{1,3} \neq \frac{1}{3}(\ln Y_1^* + \ln Y_2^* + \ln Y_3^*)$ due to Jensen's inequality. Mariano and Murasawa (2003, 2010), in a similar state-space model, document this nonlinear aggregation problem and suggest redefining the disaggregated data as the geometric mean (instead of the arithmetic mean) of the disaggregated data such that $\ln \overline{Y}_{1,3} = \frac{1}{3}(\ln Y_1^{**} + \ln Y_2^{**} + \ln Y_3^{**})$, where $\{\ln Y_t^{**}\}_{t=1}^T$ are used as a component series in the VAR system. Under this definition the disaggregated data cannot be interpreted as the calendar monthly data. They only bear a statistical interpretation such that the geometric average of latent $\ln Y_1^{**}$, $\ln Y_2^{**}$, $\ln Y_3^{**}$ equals to the observed $\ln \overline{Y}_{1,3}$. Camacho and Perez-Quiros (2010) argue that the approximation error is almost negligible if monthly changes are small and the geometric averaging works well in practice.

2.7 An application

In this section, we applied our approach to a structural VAR model to study the dynamic effects of monetary policy shocks. Note that the current model is essentially a reduced-form VAR with missing disaggregated data. Once we have decoded the latent variables and estimated the model parameters, the model is treated in the same way as a standard reduced-form VAR model. To convert the reduced form to a structural form, economic constraints must be imposed to identify the structural shocks. Our method is especially ideal for structural models with short run identification constraints.

Christiano et al. (1998) propose a block diagonal recursiveness assumption to identify monetary policy shocks. The variables in the VAR system are classified into three groups. The variables in the first group are major economic indicators. The second group consists of only

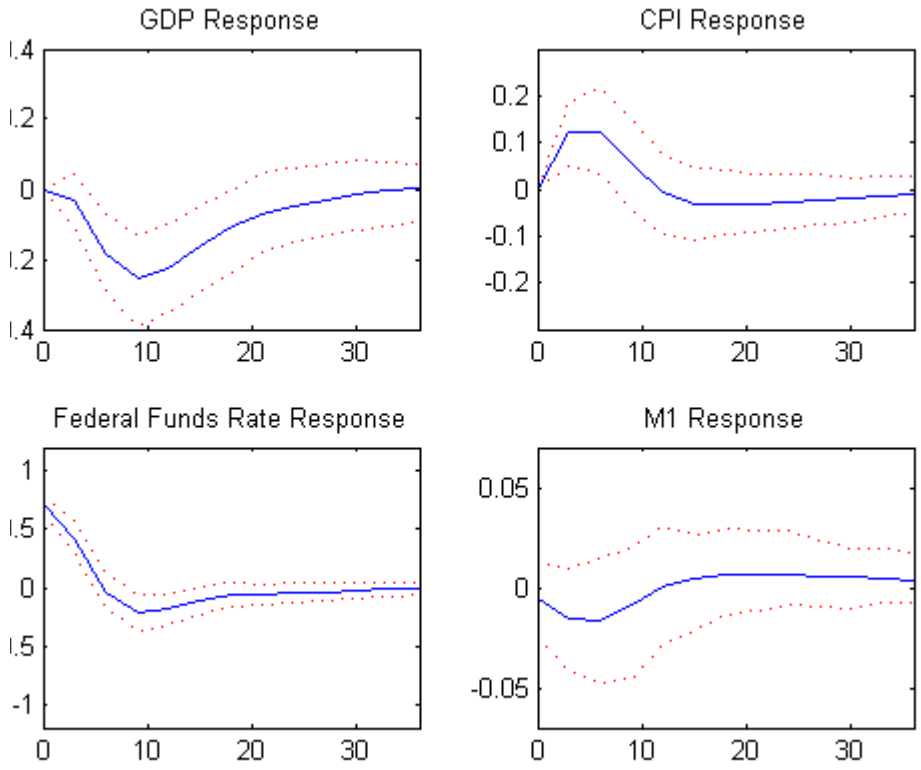
one monetary policy instrument, whose innovations reflect monetary policy shocks. Typically this variable is the federal funds rate (FF). The third group mostly include variables reflecting intermediate monetary goals. The identification assumption is that the policy instrument has no contemporaneous effect on the variables in the first group, and the third group has no contemporaneous effect on the previous two groups. This assumption enables a partial identification of monetary shocks by Cholesky decomposition, leaving other shocks unidentified.

We put GDP and the CPI in the first group, FF as the policy instrument and the money stock M1 in the third group. Monthly data of the CPI, FF and M1 are available while GDP data are quarterly. The sample period is chosen as 1974:01 to 2006:12 since there might be structural breaks before and after that time interval. Data are Hodrick-Prescott filtered for we are interested in the cyclical component of the data. Despite the common opinion that many macro-variables may contain unit roots, at least in the above sample period we did not find strong evidence against stationarity for the detrended four series. Both the ADF and Phillips-Perron tests decisively reject the null of unit roots. So we put variables in level to the model. As for the choice of lag length, there is a tradeoff between richer dynamics with more lags and unreliable estimation with more parameters. In a four-variable VAR, an additional lag means 16 more parameters. We have 396 monthly observations (132 quarterly observations). The sample size does not permit us to include many lags. The results reported in Figure 2.1 and 2.2 correspond to a quarterly VAR with 2 lags (42 parameters) and monthly model with 6 lags (106 parameters) respectively. Results robustness is checked by varying quarterly VAR lags from 1 to 6 and monthly model lags from 2 to 8. The dynamic patterns of the response functions have little change, though in the higher order system the dynamic response curves exhibit more rugs and oscillations.

Figure 2.1 and 2.2 are not directly comparable in that the short-run identification constraints, though in the same format, should be interpreted differently. In the quarterly data VAR, to identify the monetary shocks we require monetary shocks have no contemporaneous effects on the output and price in a *quarter*, while in the varied frequency model we only require no such effect in a *month*. Clearly, the latter imposes a weaker identification assumption on the contemporaneous effects, so there are grounds for believing Figure 2.2 presents a more reliable

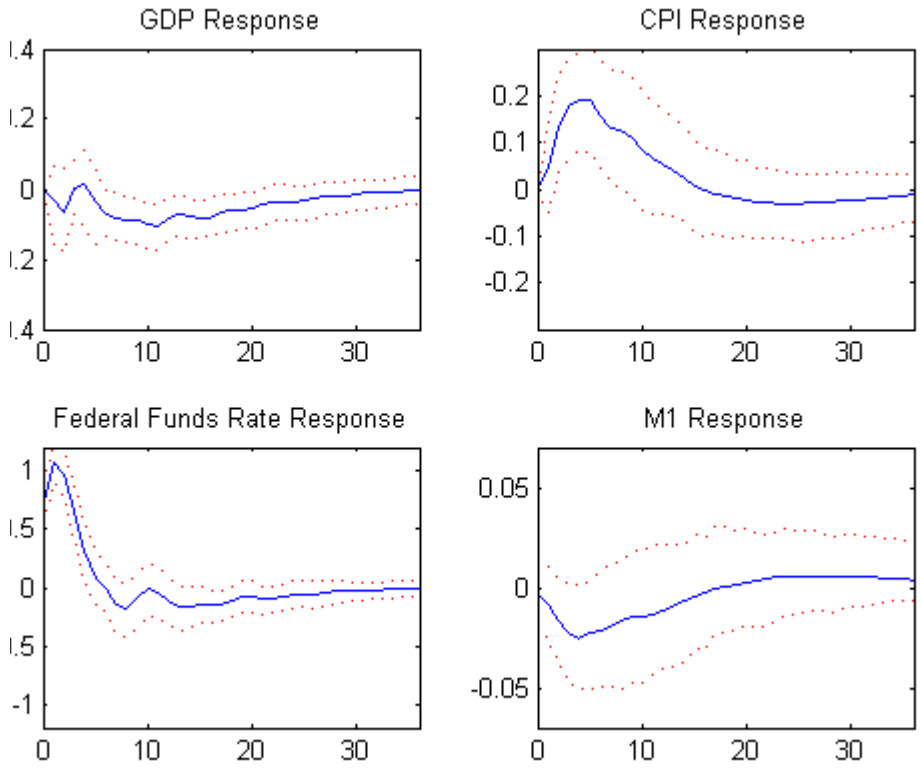
dynamic picture on how the economy responds to monetary shocks. This is a major advantage of the varied frequency model with short-run economic identification constraints. In addition, the model makes full use of available data of different frequencies. This is another reason we are in favor of the varied frequency model.

As for the dynamic patterns revealed in the Figure 2.1 and 2.2, after a contractionary monetary shock, GDP and M1 react negatively as expected, but the CPI rises steadily, a phenomenon long documented in the literature as the price puzzle (Sims, 1992; Eichenbaum, 1992). Christiano et al. (1998) suggest including an index of sensitive commodity prices can resolve the anomaly, a topic beyond the scope of the current paper.



The VAR contains two lags. One key identification constraint is monetary shocks have no contemporaneous effects on the output and price in a *quarter*. The solid line plots the posterior mean of the impulse-response function and the dotted lines are the 95% HPD credible bands. The results are obtained from a Gibbs sampler of 200000 draws with the first half of draws burned in.

Figure 2.1 Dynamic effects of monetary policy shocks using the quarterly data VAR model



The VAR contains six lags with monthly data except for quarterly GDP. One key identification constraint is monetary shocks have no contemporaneous effects on the output and price in a *month*. The solid line plots the posterior mean of the impulse-response function and the dotted lines are the 95% HPD credible bands. The results are obtained from a Gibbs sampler of 200000 draws with the first half of draws burned in.

Figure 2.2 Dynamic effects of monetary policy shocks using the varied frequency VAR model

2.8 Conclusion

The structural VAR model has been fruitfully applied in macroeconomics to unveil the dynamic paths of economic variables responding to nominal or real shocks. Such an analysis involves two stages in general. The first stage is to estimate a reduced-form VAR model and invert to its moving average representation. In the second stage, restrictions are imposed to identify structural shocks so as to conduct impulse response analysis. The second stage embodies scientific insights of the macroeconomists and is undoubtedly the core of the VAR analysis, while the first stage is largely statistical. However, a better estimate of the reduced-form VAR model translates to a more accurate impulse-response curve, and thus presents a more transparent picture of dynamic relations in the macroeconomy. The varied frequency VAR model only operates on the first stage. Varied frequency data are reconciled neither in an aggregated model with low frequency data nor in a disaggregated model with interpolated data. Instead, data of different frequencies coexist in the same model and make their due contributions to the parameter estimation. Technical details aside, the approach just provides a better estimate of the model, leaving intact the economic insights that hallmark the structural VAR analysis.

CHAPTER 3. LINEAR REGRESSION USING BOTH TEMPORALLY AGGREGATED AND TEMPORALLY DISAGGREGATED DATA: REVISITED

3.1 Introduction

Incomplete data is a common problem in applied economics. In the regression analysis, there are occasions when complete data of many relevant regressors are collected, but data on one or more key covariates are aggregated by group, by region, by time, and so on. To make the best use of the available data and minimize information loss, we hope to use both aggregated and disaggregated data in a regression. The conventional wisdom is a two-equation least squares (LS) in which the first regression imputes the unobservable disaggregated data, and then the imputed covariate data are used in the second regression. Three decades ago, with the same title, Hsiao (1979) and Palm and Nijman (1982) consider the maximum likelihood (ML) estimation of an aggregated covariate data (ACD) model in which data are measured at different temporal frequencies. However, this approach receives little attention in the subsequent empirical work. Perhaps part of the reason is that Palm and Nijman find that the likelihood function cannot be factorized into two separable parts as suggested by Hsiao. Palm and Nijman conclude that the computational advantage of ML is lost in the ACD model.

This paper revisits the model in Hsiao (1979). One contribution of this paper is that the likelihood function is found to be separable by suitable reparameterization if one instrument corresponds to one endogenous regressor. In that case, an analytic full-information ML estimator does exist and can be obtained by two auxiliary regressions. That implies an efficient estimator can be secured without computational barriers, and thus overshadows the LS imputation approach.

Our idea of likelihood separability is inspired by the statistics literature on missing data. In contrast to the standard missing data problem where a fraction of observations are unavailable, data aggregation blinds all individual-level observations, leaving relatively few aggregated data. However, once the correlation structure of aggregated and disaggregated values is properly addressed, the ACD model bears many similarities to the missing data problem. Anderson (1957), in the context of missing multivariate normal variates, raises the important idea of factoring the likelihood function into two parts, each of which can be maximized analytically. Gourieroux and Monfort (1981) extended that method to regression models with missing covariate data. In this paper, we follow this track and extend the idea of likelihood separability to the ACD model.

We are aware that not every ACD model specification satisfies the likelihood separability conditions. Furthermore, the practitioners may have their own models while at the same time aggregated covariates are involved. In that case, the likelihood function has to be maximized numerically in general. As an alternative to numerical ML, we propose a competing Bayesian approach implemented by the Gibbs sampler, which is another contribution of this paper. For models without analytic solutions, our Monte Carlo study shows that the Bayesian estimator is more robust and less sensitive to the initial values.

Our third contribution is a critique on LS imputation approaches applied to the ACD model. The asymptotics of LS-type estimators have been extensively discussed in the literature. Gourieroux and Monfort (1981) provide asymptotic comparisons of a variety of LS estimators for the missing data regression. In Hsiao (1979) and Palm and Nijman (1982) there are also comparisons the relative efficiency of ML estimator with LS-type estimators for the ACD regression. In addition, simulation-based multiple imputation strategies (see Rubin, 1987; Schafer, 1997; Allison, 2000) can also be used to compute the asymptotic standard error of those estimators. However, for a ACD regression model with endogenity problems, some LS-type estimators are not consistent, and some consistent estimators discard apparent information. Those drawbacks are overcome by the ML and Bayesian estimators, which is the main reason we do not recommend the usage of LS imputation.

The model in Hsiao (1979) is originally designed for temporal aggregation, which is com-

monly found in macroeconomic and financial data. Temporal aggregation and mixed sampling frequencies regression can be appropriately tackled by time series techniques. Geweke (1978), Ghysels et al. (2006) and Andreou et al. (2010) develop corresponding models and estimation techniques. Revisiting Hsiao (1979), we feel that his model might be most suitable for aggregation problems encountered in applied microeconomics. We illustrate the potential applications of the ACD regression with three examples.

Example 3.1. *We want to evaluate the impact of the Low-Income Home Energy Assistance Program (LIHEAP) on the subsequent energy expenditures of its recipients. The LIHEAP grant is a one-time payment for the winter season, whereas the gas or electricity is always billed monthly. Although we can aggregate the monthly bills as well and conduct an analysis at the seasonal level, we lose the monthly information contained in the dependent variable,and other covariates such as monthly income and weather. Now consider the ACD regression: monthly usage of the grant (in consumption or saving) is latent, up to the individual choice and summing up to the observable total amount. If we can impute the latent monthly grant usage, we will know what proportion of the grant contributes to monthly energy expenditures.*

Example 3.2. *Occupational Outlook Handbook (Bureau of Labor Statistics, U.S. Department of Labor, 2010) predicts that veterinarians will increase by 33% over the 2008–18 decade, much faster than the average for all occupations. Suppose we want to study whether the fast growth of the cat (pet) population pushes up the demand for veterinary services. We searched the database for public use and found that the veterinarian data, along with many other covariates, of each county are available, while the pet population is only recorded for each state, hence the ACD.*

Example 3.3. *In development economics, we might be interested in the calorie-income elasticity in poor countries. The calorie intake is an individual measure, varying among men, women and children. However, the observable household income is likely to be redistributed within the family, and thus the real individual income is a latent regressor to the researcher.*

There are many practical reasons the covariate data are aggregated. In Example 1 and 3, by the nature of the variable, the disaggregated values are never observed. Example 2 illustrates that data collection difficulties, confidentiality of the personal information, and grouping during

the dataset construction often lead to aggregated variables. This is especially true for the publicly accessible dataset.

The rest of the paper is organized as follows. Section 3.2 presents the ACD model. Section 3.3 derives the full-information likelihood function and discusses conditions for separability. Section 3.4 proposes a competing Bayesian estimator using the Gibbs sampler. Section 3.5 briefly reviews the traditional least squares based solutions. Section 3.6 compares various estimators by Monte Carlo experiments. Section 3.7 extends the model to multiple aggregated covariates, imbalanced aggregation, as well as partial aggregation. Section 3.8 concludes the paper.

3.2 The ACD model

We follow the model and notation in Hsiao (1979) and Palm and Nijman (1982), but add a richer set of regressors for better control of partial effects.

The ACD model consists of the following equations:

$$y_{t,i} = x_{t,i}\beta + \mathbf{w}_{t,i}\delta + u_{t,i} \tag{3.1}$$

$$x_{t,i} = \mathbf{z}_{t,i}\alpha + \mathbf{w}_{t,i}\gamma + v_{t,i} \tag{3.2}$$

$$\overline{x}_t = \sum_{i=1}^{n} x_{t,i} \tag{3.3}$$

where

$$\begin{pmatrix} u_{t,i} \\ v_{t,i} \end{pmatrix} \sim i.i.d.N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}\right], t = 1, ..., T; i = 1, ..., n.$$

Exogenous explanatory variables (row vectors) $\mathbf{w}_{t,i}, \mathbf{z}_{t,i}$ are uncorrelated with disturbance terms, and parameters δ, α, γ are column vectors.

Eq. (3.1) is the primary regression model and β is the parameter of main interest. However, the key covariate data $x_{t,i}$ are unavailable at the disaggregated level, with only aggregated values $\overline{x}_t$ being observed. Other variables $y_{t,i}, \mathbf{w}_{t,i}, \mathbf{z}_{t,i}$ all have complete data. As in Hsiao (1979), the subscript (t, i) originally refers to the i^{th} observation in the year t. That is, semiannual (or quarterly, monthly) data are aggregated into annual data. In more general settings, we

may interpret t as the group index, and i as the i^{th} member in that group. Data of individual members in a group are aggregated. For instance, in Example 2, $x_{t,i}$ refers to the latent pet population in the county i of the state t, but only the state-level population $\overline{x}_t$ is observed.

Eq. (3.2) is the imputation model for the unobserved $x_{t,i}$. The choice of variables for imputation was discussed in Schafer (1997) and Van Buuren et al. (1999). They suggested that covariates in the main regression (i.e. $\mathbf{w}_{t,i}$) should be included, and factors related to missing mechanisms and with substantial explanatory power over $x_{t,i}$ can also be included, which are captured in $\mathbf{z}_{t,i}$. For instance, in Example 1, in addition to $x_{t,i}$ (the latent grant usage) which explains the monthly energy bill, a plausible set of regressors in $\mathbf{w}_{t,i}$ may include the outdoor temperature, household income, family and room size, the age indicator variable, etc. To impute $x_{t,i}$, we may add all variables in $\mathbf{w}_{t,i}$ and monthly saving-to-income ratio as $\mathbf{z}_{t,i}$.

Of course, the data aggregation model *per se* does not require the appearance of $\mathbf{w}_{t,i}$ in both Eq. (3.1) and (3.2). Even if some or none of the variables in $\mathbf{w}_{t,i}$ are included in Eq. (3.2), the model is still estimable by both ML and Bayesian methods. However, the separability of the likelihood and the closed-form solution requires the presence of $\mathbf{w}_{t,i}$ in Eq. (3.2).

The relationship of disturbance terms across two equations determines the role of $x_{t,i}$ in Eq. (3.1). If we allow the possibility that some unmodelled factors (maybe because of data collection difficulties) can affect both $x_{t,i}$ and $y_{t,i}$, then $u_{t,i}$ and $v_{t,i}$ are correlated. In that case, $x_{t,i}$ is an endogenous regressor in Eq. (3.1). On the other hand, by the exogeneity assumption on $\mathbf{z}_{t,i}$, it satisfies all the requirements of a valid instrument. Of course, the data aggregation model *per se* is not necessarily associated with endogeneity. Maybe $\mathbf{z}_{t,i}$ is included solely because it can better explain and impute the missing $x_{t,i}$. However, as we will see below, if we do have one instrument corresponding to one endogenous regressor, we allow the separability of the likelihood function.

Though disturbances across equations are allowed to be correlated, throughout this paper we assume no serial correlation. That is, changing subscript either t or i will lead to uncorrelated disturbances. If we had long time series aggregated at varied frequencies, it would be more appropriate to infer the dependency structure of disaggregated series from observed aggregated series. However, in microeconomic applications, the aggregation is often at the geographic or

individual level as in Example 2 and 3, and the dependency structure is not obvious. Example 1 does involve temporal aggregation, but there are only 4 or 5 months in winter, so it is harder to model their dependency.

3.3 Maximum likelihood estimation

3.3.1 Joint likelihood

Although the ACD model can be estimated by LS procedures, this approach is not efficient (see Palm and Nijman (1982) and Gourieroux and Monfort (1981) for discussion). An efficient estimator of $\theta \equiv \left(\alpha,\beta,\gamma,\delta,\sigma_u^2,\sigma_v^2,\sigma_{uv}\right)$ can be obtained by making full use of the information conveyed by the observed data, maximizing the joint likelihood

$$\ln L(\theta) = \sum_{t=1}^{T} \ln f\left(y_{t,1},...,y_{t,n},\overline{x}_t\right)$$

conditional on the exogenous regressors $\mathbf{w}_{t,i},\mathbf{z}_{t,i}$.

Hsiao (1979) and Palm and Nijman (1982) derive the likelihood for the case of $n = 2$, that is, $\sum_{t=1}^{T} \ln f\left(y_{t,1},y_{t,2},\overline{x}_t\right)$. Hsiao (1979) first introduced $x_{t,1}$ into the likelihood and then integrated it out: $f\left(y_{t,1},y_{t,2},\overline{x}_t\right) = \int f\left(y_{t,1},y_{t,2}\left|\overline{x}_t,x_{t,1}\right.\right) \cdot f\left(\overline{x}_t,x_{t,1}\right) dx_{t,1}$. Palm and Nijman (1982) derived an equivalent form of the likelihood $f\left(y_{t,1}+y_{t,2},y_{t,1}-y_{t,2},\overline{x}_t\right)$ by integration with respect to $x_{t,1}$.

In fact, a shortcut to obtain the joint likelihood is by manipulation of Eq. (3.1) to (3.3).

First of all, define the following symbols:

$$\begin{aligned}
\overline{y}_t &= \sum_{i=1}^{n} y_{t,i},\ \overline{\mathbf{z}}_t = \sum_{i=1}^{n} \mathbf{z}_{t,i},\ \overline{\mathbf{w}}_t = \sum_{i=1}^{n} \mathbf{w}_{t,i}, \\
\mathbf{y}_t &= \begin{pmatrix} y_{t,1} \\ ... \\ y_{t,n} \end{pmatrix},\ \mathbf{x}_t = \begin{pmatrix} x_{t,1} \\ ... \\ x_{t,n} \end{pmatrix},\ \mathbf{z}_t = \begin{pmatrix} \mathbf{z}_{t,1} \\ ... \\ \mathbf{z}_{t,n} \end{pmatrix},\ \mathbf{w}_t = \begin{pmatrix} \mathbf{w}_{t,1} \\ ... \\ \mathbf{w}_{t,n} \end{pmatrix}.
\end{aligned}$$

Plugging Eq. (3.2) into Eq. (3.1) and (3.3), we have

$$y_{t,i} = \mathbf{z}_{t,i}\alpha\beta + \mathbf{w}_{t,i}\left(\beta\gamma+\delta\right) + \left(\beta v_{t,i} + u_{t,i}\right),$$

$$\overline{x}_t = \overline{\mathbf{z}}_t\alpha + \overline{\mathbf{w}}_\mathbf{t}\gamma + \left(v_{t,1} + ... + v_{t,n}\right).$$

Since $(v_{t,1}, ..., v_{t,n}, u_{t,1}, ..., u_{t,n})$ can be viewed as $2n$ dimensional multivariate normal, their $n+1$ dimensional (mean-adjusted) linear combinations $(y_{t,1}, ..., y_{t,n}, \overline{x}_t)$ are also multivariate normal, and we have

$$\begin{pmatrix} \mathbf{y}_t \\ \overline{x}_t \end{pmatrix} \sim N \left\{ \begin{bmatrix} \mathbf{z}_t\alpha\beta + \mathbf{w}_t(\beta\gamma + \delta) \\ \overline{\mathbf{z}}_t\alpha + \overline{\mathbf{w}}_\mathbf{t}\gamma \end{bmatrix}, \begin{bmatrix} \left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)\mathbf{I}_n & \left(\beta\sigma_v^2 + \sigma_{uv}\right)\iota_n \\ \left(\beta\sigma_v^2 + \sigma_{uv}\right)\iota_n' & n\sigma_v^2 \end{bmatrix} \right\},$$

where $\mathbf{I}_n$ is the identity matrix, and ι_n is a column vector of ones.

If we decompose the joint multivariate normal density into

$$f(\mathbf{y}_t, \overline{x}_t) = f(\mathbf{y}_t | \overline{x}_t) \cdot f(\overline{x}_t),$$

we will arrive at expression (11) on p.246 in Hsiao (1979) where $f(\mathbf{y}_t | \overline{x}_t)$ is termed L_1 and $f(\overline{x}_t)$ termed L_2. Palm and Nijman (1982) had the same in expression (4) on p.335 of their paper.

3.3.2 Separability of likelihood

The likelihood function $L(\theta)$ is separable if it can be factorized as

$$L(\theta) = L_1(\theta_1) \cdot L_2(\theta_2),$$

where (θ_1, θ_2) is a partition of θ.

A separable likelihood function has a computational advantage in that maximization with respect to θ can be performed through $\max_{\theta_1} L_1(\theta_1)$ and $\max_{\theta_2} L_2(\theta_2)$ respectively. Moreover, Anderson (1957) discovered that those two maximizations may have analytic solutions for some (but not all) types of missing multivariate normal variates.

For the ACD model, Palm and Nijman (1982) pointed out that the L_1 and L_2 in Hsiao (1979) are not separable. However, there are two useful special cases in which separability does exist. To find the separable form, we first factorize the joint density in the other order:

$$f(\mathbf{y}_t, \overline{x}_t) = f(\mathbf{y}_t) \cdot f(\overline{x}_t | \mathbf{y}_t),$$

and then we reparameterize the model and construct the partition.

The first case is when $z_{t,i}$ is a scalar variable (so is α), and no restrictions are imposed on σ_{uv}.

Then we have

$$f(\mathbf{y}_t) = \phi(\mathbf{y}_t; \mathbf{z}_t \cdot A + \mathbf{w}_t \cdot \mathbf{B}, C \cdot \mathbf{I}_n),$$

$$f(\overline{x}_t | \mathbf{y}_t) = \phi(\overline{x}_t; \overline{z}_t \cdot D + \overline{\mathbf{w}}_t \cdot \mathbf{E} + \overline{y}_t \cdot F, G),$$

where $\phi(\mathbf{y}; \mu, \mathbf{\Sigma})$ is the density of $N(\mu, \mathbf{\Sigma})$ evaluated at $\mathbf{y}$, and

$$\begin{aligned}
A &= \alpha\beta, \\
\mathbf{B} &= \beta\gamma + \delta, \\
C &= \beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}, \\
D &= \alpha - \left(\beta\sigma_v^2 + \sigma_{uv}\right)\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}\alpha\beta, \\
\mathbf{E} &= \gamma - \left(\beta\sigma_v^2 + \sigma_{uv}\right)\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}(\beta\gamma + \delta), \\
F &= \left(\beta\sigma_v^2 + \sigma_{uv}\right)\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}, \\
G &= n\sigma_v^2 - n\left(\beta\sigma_v^2 + \sigma_{uv}\right)^2\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}.
\end{aligned}$$

The derivation is straightforward in that $f(\mathbf{y}_t)$ and $f(\overline{x}_t | \mathbf{y}_t)$ are simply the marginal and conditional density of the multivariate normal distribution. However, the result implies that the likelihood function have a separable form with respect to the new parameters, which can be partitioned as $(A, \mathbf{B}, C), (D, \mathbf{E}, F, G)$.

Furthermore, note that

$$\max_{A,\mathbf{B},C} \sum_{t=1}^{T} \ln f(\mathbf{y}_t)$$

is equivalent to the ML estimation of the linear regression

$$y_{t,i} = z_{t,i} \cdot A + \mathbf{w}_{t,i} \cdot \mathbf{B} + \varepsilon_{t,i},$$

where $\varepsilon_{t,i} \sim N(0, C)$. The analytic ML estimator is given by

$$\begin{pmatrix} \widehat{A} \\ \widehat{\mathbf{B}} \end{pmatrix} = \left[\sum_{t=1}^{T}\sum_{i=1}^{n} (z_{t,i}, \mathbf{w}_{t,i})' (z_{t,i}, \mathbf{w}_{t,i})\right]^{-1} \left[\sum_{t=1}^{T}\sum_{i=1}^{n} (z_{t,i}, \mathbf{w}_{t,i})' y_{t,i}\right],$$

$$\widehat{C} = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(y_{t,i} - z_{t,i}\widehat{A} - \mathbf{w}_{t,i}\widehat{\mathbf{B}}\right)^2.$$

Similarly,

$$\max_{D,\mathbf{E},F,G} \sum_{t=1}^{T} \ln f\left(\overline{x}_t \left| \mathbf{y}_t \right.\right)$$

is equivalent to the ML estimation of the linear regression

$$\overline{x}_t = \overline{z}_t \cdot D + \overline{\mathbf{w}}_t \cdot \mathbf{E} + \overline{y}_t \cdot F + \eta_t,$$

where $\eta_t \sim N\left(0, G\right)$. The ML estimator is given by

$$\begin{pmatrix} \widehat{D} \\ \widehat{\mathbf{E}} \\ \widehat{F} \end{pmatrix} = \left[\sum_{t=1}^{T} \left(\overline{z}_t, \overline{\mathbf{w}}_t, \overline{y}_t\right)' \left(\overline{z}_t, \overline{\mathbf{w}}_t, \overline{y}_t\right)\right]^{-1} \left[\sum_{t=1}^{T} \left(\overline{z}_t, \overline{\mathbf{w}}_t, \overline{y}_t\right)' \overline{x}_t\right],$$

$$\widehat{G} = \frac{1}{T} \sum_{t=1}^{T} \left(\overline{x}_t - \overline{z}_t \widehat{D} - \overline{\mathbf{w}}_t \widehat{\mathbf{E}} - \overline{y}_t \widehat{F}\right)^2.$$

Finally, since the ML estimator is invariant to the one-to-one reparameterization, the full-information ML estimator for $\left(\alpha,\beta,\gamma,\delta,\sigma_u^2,\sigma_v^2,\sigma_{uv}\right)$ is related to $\left(A,\mathbf{B},C,D,\mathbf{E},F,G\right)$ by the following formula:

$$\begin{aligned}
\alpha &= D + AF, \\
\beta &= \frac{A}{D + AF}, \\
\gamma &= \mathbf{E} + \mathbf{B}F, \\
\delta &= \frac{\mathbf{B}D - A\mathbf{E}}{D + AF}, \\
\sigma_u^2 &= \frac{A^2 G + nCD^2}{n\left(D + AF\right)^2}, \\
\sigma_v^2 &= CF^2 + \frac{1}{n}G, \\
\sigma_{uv} &= \frac{nCDF - AG}{n\left(D + AF\right)}.
\end{aligned}$$

There is one issue to clarify. Note that the covariance matrix of $u_{t,i}, v_{t,i}$ must be positive definite, which imposes inequality constraints $\sigma_u^2 > 0$, $\sigma_v^2 > 0$, and $\sigma_u^2\sigma_v^2 - \sigma_{uv}^2 > 0$. The two auxiliary regressions yield $\widehat{C} > 0$, $\widehat{G} > 0$ by construction, so we immediately have $\widehat{\sigma_u^2} > 0$, $\widehat{\sigma_v^2} > 0$. Furthermore, a little algebra reveals $\widehat{\sigma_u^2}\widehat{\sigma_v^2} - \widehat{\sigma_{uv}^2} > 0$. In a word, the above procedure guarantees that inequality constraints are automatically satisfied.

The second case is that $z_{t,i}$ does not exist, and σ_{uv} is restricted to zero.

Then we have

$$f(\mathbf{y}_t) = \phi(\mathbf{y}_t; \mathbf{w}_t \cdot \mathbf{B}, C \cdot \mathbf{I}_n),$$
$$f(\overline{x}_t | \mathbf{y}_t) = \phi(\overline{x}_t; \overline{\mathbf{w}}_t \cdot \mathbf{E} + \overline{y}_t \cdot F, G),$$

where

$$\begin{aligned}
\mathbf{B} &= \beta\gamma + \delta, \\
C &= \beta^2\sigma_v^2 + \sigma_u^2, \\
\mathbf{E} &= \gamma - \left(\beta\sigma_v^2\right)\left(\beta^2\sigma_v^2 + \sigma_u^2\right)^{-1}(\beta\gamma + \delta), \\
F &= \left(\beta\sigma_v^2\right)\left(\beta^2\sigma_v^2 + \sigma_u^2\right)^{-1}, \\
G &= n\sigma_v^2 - n\left(\beta\sigma_v^2\right)^2\left(\beta^2\sigma_v^2 + \sigma_u^2\right)^{-1}.
\end{aligned}$$

The separability of the likelihood implies $\widehat{\mathbf{B}}, \widehat{C}$ can be obtained from the linear regression of $y_{t,i}$ on $\mathbf{w}_{t,i}$, $\widehat{\mathbf{E}}, \widehat{F}, \widehat{G}$ can be obtained from the regression of $\overline{x}_t$ on $\overline{\mathbf{w}}_t$, $\overline{y}_t$. Then the full-information ML estimator of $\left(\beta, \gamma, \delta, \sigma_u^2, \sigma_v^2\right)$ can be solved with the following closed form:

$$\begin{aligned}
\beta &= \frac{nCF}{nCF^2 + G}, \\
\gamma &= \mathbf{E} + \mathbf{B}F, \\
\delta &= \frac{\mathbf{B}G - nC\mathbf{E}F}{nCF^2 + G}, \\
\sigma_u^2 &= \frac{CG}{nCF^2 + G}, \\
\sigma_v^2 &= CF^2 + \frac{1}{n}G.
\end{aligned}$$

The above two cases deserve some remarks.

First, if the ACD model specification does not belong to the two special cases, it does not mean the model is not estimable by ML. As long as the model is identifiable, the likelihood can always be maximized by numerical procedures. However, separability of the likelihood offers a computational advantage — it is even less costly than the imputed value two-step estimator. Note that both point estimators are computed from two OLS regressions, but the standard

error adjustment for the two-step estimator is not straightforward, while standard errors for the ML estimator can be computed by the Delta method.

Second, as far as an applied problem is concerned, the two special cases are not so restrictive as they might seem. Separability of the likelihood can be achieved if we reasonably redesign the model in use. Case 1 requires one instrument variable $z_{t,i}$ corresponds to one endogenous aggregated regressor $x_{t,i}$. Suppose the goal is to impute the aggregated $x_{t,i}$, but the endogeneity is not of major concern. By allowing the possibility of non-zero σ_{uv}, we gain, in addition to $\mathbf{w}_{t,i}$, another variable $z_{t,i}$ which lends explanatory power to impute $x_{t,i}$. If more than one such additional variables are available, we might consider extracting the first principal component of them. In that fashion we retain most of the explanatory power for imputation and meanwhile save computational costs. Case 2 is suitable when $x_{t,i}$ is not endogenous, but the only set of regressors $\mathbf{w}_{t,i}$ appearing in both Eq. (3.1) and (3.2) seems restrictive. If we have different covariates for the two equations in mind, we can take the union of the regressors to form $\mathbf{w}_{t,i}$.

Lastly, if $z_{t,i}$ does not exist, and we allow $\sigma_{uv} \neq 0$, the model is not identified, and thus should be avoided.

3.4 Bayesian estimator

If either special case is satisfied, the analytic ML estimator is our first choice of estimating the ACD model for the sake of efficiency and computability. However, we are aware that there are circumstances when i) we cannot revise our model catering to the special cases, and the numerical ML does not perform satisfactorily; ii) we have prior information or beliefs on the parameter values or restrictions on the parameters; or iii) we are primarily estimating other models, meanwhile some data are aggregated. In these situations, the likelihood might be difficult to formulate and maximize numerically. We therefore propose a competing Bayesian approach, in which the joint posteriors of the latent covariate as well as other parameters of uncertainty are simulated using the Gibbs sampler.

Though frequentist and Bayesian inferences handle parameter uncertainty differently, both make use of the sampling distribution, or the "likelihood function" when the sampling dis-

tribution is viewed as a function of model parameters. From the Bayesian perspective, the posteriors are proportional to the priors times the likelihood. When the priors are diffuse, the posteriors virtually inherit the shape of the likelihood function. A Bayesian "point estimator", under the all-or-nothing decision rules, can be the mode of the posteriors, which coincides with ML estimator since ML seeks the peak of the likelihood function. In this sense, the classic and Bayesian inferences are comparable since the information contained in the sampling distribution is the same. A pragmatic difference is that numerical ML has limited ability to locate the peak if the starting values are not carefully specified, while the Bayesian MCMC simulation with flat prior can reliably recover the entire shape of the likelihood.

The Gibbs sampler cycles through the full conditional posteriors (each variable or variables block conditional on other variables as well as the data). In latent variable models, posterior conditionals for model parameters would be of standard form if the latent variable were known. The key step is to specify the posterior conditionals for the latent variable.

Let us first define the following symbols:

$$\psi = (\beta, \delta', \alpha', \gamma')', \ \mathbf{X}_{ti} = \begin{bmatrix} (x_{t,i}, \mathbf{w}_{t,i}) & \mathbf{0} \\ \mathbf{0} & (\mathbf{z}_{t,i}, \mathbf{w}_{t,i}) \end{bmatrix},$$

$$\mathbf{Y}_{t,i} = \begin{pmatrix} y_{t,i} \\ x_{t,i} \end{pmatrix}, \ \mathbf{\Sigma} = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}.$$

Conjugate proper priors are specified as:

$$\psi \sim N\left(\underline{\mu}, \underline{\mathbf{V}}\right),$$
$$\mathbf{\Sigma}^{-1} \sim Wishart\left(\underline{\mathbf{\Omega}}, \underline{\nu}\right).$$

The hyperparameters $\underline{\mu}, \underline{\mathbf{V}}, \underline{\mathbf{\Omega}}, \underline{\nu}$ can be set to contain little information, so that posteriors are mostly learned from the likelihood function.

Conditional on the latent $\{x_{t,i}\}$, it is a standard seemingly unrelated regression (SUR) model. The full posterior conditionals are (refer to the textbook Koop et al. (2007) for a

derivation):

$$\psi \,|\cdot \sim N\left(\mathbf{Dd}, \mathbf{D}\right),$$
$$\boldsymbol{\Sigma}^{-1} \,|\cdot \sim Wishart\left(\overline{\boldsymbol{\Omega}}, \overline{\nu}\right),$$

where

$$\begin{aligned}
\mathbf{D} &= \left(\sum_{t=1}^{T}\sum_{i=1}^{n} \mathbf{X}_{ti}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_{ti} + \underline{\mathbf{V}}^{-1}\right)^{-1}, \\
\mathbf{d} &= \sum_{t=1}^{T}\sum_{i=1}^{n} \mathbf{X}_{ti}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}_{t,i} + \underline{\mathbf{V}}^{-1}\underline{\mu}, \\
\overline{\boldsymbol{\Omega}} &= \left[\underline{\boldsymbol{\Omega}}^{-1} + \sum_{t=1}^{T}\sum_{i=1}^{n} \left(\mathbf{Y}_{t,i} - \mathbf{X}_{ti}\psi\right)\left(\mathbf{Y}_{t,i} - \mathbf{X}_{ti}\psi\right)'\right]^{-1}, \\
\overline{\nu} &= \underline{\nu} + nT.
\end{aligned}$$

To derive the full posterior conditional distribution for the latent $\{x_{t,i}\}$, we first introduce a proposition on restricted multivariate normal distribution. Fraser (1951) solved for the general form of n-dimension distribution subject to k $(k < n)$ linear constraints by transforming the linear space, but his procedure is descriptive and no explicit distributional forms are given. However, for the purpose of this paper, we only need to solve for a special case — n originally uncorrelated normal variates subject to one aggregation constraint. Explicit solutions are provided in the following proposition (See the appendix for a proof).

Proposition 3.1. *Let* $\mathbf{x} = (x_1, ..., x_n)'$ *be a multivariate normal random vector with zero correlations.* $x_i \sim N\left(\mu_i, \sigma^2\right), i = 1, ..., n$. *Conditional on the aggregation constraint:* $\sum_{i=1}^{n} x_i = \overline{x}$ *where* $\overline{x}$ *is fixed, we have*

$$\mathbf{x}_{-n} \,|\overline{x} \sim N\left[\mu_{-n} + \frac{1}{n}\left(\overline{x} - \sum_{i=1}^{n}\mu_i\right)\iota_{n-1}, \sigma^2\left(\mathbf{I}_{n-1} - \frac{1}{n}\iota_{n-1}\iota_{n-1}'\right)\right],$$

where $\mathbf{x}_{-n} = (x_1, ..., x_{n-1})'$, $\mu_{-n} = (\mu_1, ..., \mu_{n-1})'$, $\mathbf{I}_{n-1}$ *is the identity matrix, and* ι_{n-1} *is a vector of ones. Moreover,* $x_n \,|\mathbf{x}_{-n}, \overline{x}$ *is degenerated, and equals to* $\overline{x} - \sum_{i=1}^{n-1} x_i$.

Proof. See appendix. □

Marginally, $x_i \left| \overline{x} \right. \sim N\left[\mu_i + \frac{1}{n}\left(\overline{x} - \sum_{i=1}^{n}\mu_i\right), \left(1 - \frac{1}{n}\right)\sigma^2\right]$ for $i = 1, ..., n$. However, only $n-1$ of them can form a multivariate normal distribution, with the remaining variable having a degenerated distribution. To sample from that restricted distribution, we first take draws from $f\left(\mathbf{x}_{-n} \left| \overline{x} \right.\right)$, and then subtract $\sum_{i=1}^{n-1} x_i$ from $\overline{x}$ to obtain x_n.

We now derive the posterior conditional distribution of the latent $\{x_{t,i}\}$ in the ACD model. The latent covariate data are uncorrelated unconditionally, but correlated conditional on the aggregation constraint. So for each group $t = 1, ...T$, we sample $(x_{t,1}, ..., x_{t,n})$ using Proposition 3.1. The distributional form is provided in the next proposition (See the appendix for a proof).

Proposition 3.2. *For every group t, the full posterior conditional $\mathbf{x}_t \left| \cdot \right.$ can be decomposed as*

$$\begin{aligned}
\mathbf{x}_{t,-n} \left| \cdot \right. &\sim N\left[\overline{\mu}_{t,-n} + \frac{1}{n}\left(\overline{x}_t - \sum_{i=1}^{n}\overline{\mu}_{t,i}\right)\iota_{n-1}, \overline{\sigma}^2\left(\mathbf{I}_{n-1} - \frac{1}{n}\iota_{n-1}\iota_{n-1}'\right)\right], \\
x_{t,n} \left| \cdot, \mathbf{x}_{t,-n} \right. &= \overline{x}_t - \sum_{i=1}^{n-1} x_{t,i},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{x}_{t,-n} &= (x_{t,1}, ..., x_{t,n-1})', \\
\overline{\mu}_{t,-n} &= \left(\overline{\mu}_{t,1}, ..., \overline{\mu}_{t,n-1}\right)', \\
\overline{\mu}_{t,i} &= \mathbf{z}_{t,i}\alpha + \mathbf{w}_{t,i}\gamma + \frac{\beta\sigma_v^2 + \sigma_{uv}}{\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}}\left[y_{t,i} - \mathbf{z}_{t,i}\alpha\beta - \mathbf{w}_{t,i}\left(\beta\gamma + \delta\right)\right], \\
\overline{\sigma}^2 &= \sigma_v^2 - \left(\beta\sigma_v^2 + \sigma_{uv}\right)^2\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}.
\end{aligned}$$

Proof. See appendix. □

The result is conformable with the "Exercise 14.19 (Missing data, 3)" in Koop et al. (2007), That exercise solves a missing data problem with a univariate regression imputation. The intuition underlying the approach is that our knowledge of missing data is updated by two pieces of information: one from the main regression equation, while the other from the imputation equation. The ACD proceeds further, since there is a third piece of information from the aggregation constraint.

The Gibbs sampler cycles through $\psi \left| \cdot \right.$, $\mathbf{\Sigma}^{-1} \left| \cdot \right.$ and $\mathbf{x}_t \left| \cdot \right.$, $t = 1, ..., T$. Once the chain converges, we obtain posterior draws from the joint distribution of $\psi, \mathbf{\Sigma}^{-1}, \{\mathbf{x}_t\}$ conditional on $\{y_{t,i}\}, \{\overline{x}_t\}$.

Note that our prior knowledge can be flexibly incorporated into the Bayesian model. For instance, if we know that parameters must belong to some set, we might use truncated priors; if the parameters are subject to equality or inequality constraints, methods in Geweke (1995) can be employed; if we take an objective Bayesian stance, we might use non-informative priors for ψ and $\mathbf{\Sigma}^{-1}$. In all those cases, sampling procedures for model parameters may change accordingly, but the essential step to sample the latent $\{x_{t,i}\}$ remains the same.

There are also circumstances when some other models are of primary interest while some data are aggregated. The Bayesian procedure is flexible enough to handle that complicity. For example, when estimating a Probit model where $\{y_{t,i}\}$ is binary, while at the same time one covariate $\{x_{t,i}\}$ is aggregated, the standard Gibbs sampler for the Probit model can still be used, with the insertion of an additional step outlined earlier to sample the latent covariate.

3.5 Least squares estimators

For completeness of estimation strategies, we outline several ways to estimate the ACD model on the basis of LS.

The first approach is an all-aggregated-data estimator. Since the parameters of primary interest are β and δ, we effectively ignore the imputation regression Eq. (3.2), but aggregate $y_{t,i}$ and $\mathbf{w}_{t,i}$ as well to regress

$$\overline{y}_t = \overline{x}_t\beta + \overline{\mathbf{w}}_t\delta + \overline{u}_t.$$

This estimator is consistent as $T \to \infty$, the asymptotic variance is n times larger than what it would be attained by regressing Eq. (3.1) if complete data were observed.

The second approach is a two-step estimator due to Dagenais (1973), which is used to address the conventional missing data problems. In the first step, we use aggregated data $\{\overline{x}_t, \overline{\mathbf{z}}_t, \overline{\mathbf{w}}_t\}$ to fit Eq. (3.2), and then use disaggregated data $\left\{\mathbf{z}_{t,i}, \mathbf{w}_{t,i}\right\}$ to predict (impute) $\{x_{t,i}\}$ as

$$\widehat{x}_{t,i} = \mathbf{z}_{t,i}\widehat{\alpha} + \mathbf{w}_{t,i}\widehat{\gamma},$$

where $\widehat{\alpha}, \widehat{\gamma}$ is the OLS estimator of regressing $\overline{x}_t$ on $\overline{\mathbf{z}}_t, \overline{\mathbf{w}}_t$.

In the second step, with $\widehat{x}_{t,i}$ in place of $x_{t,i}$, we regress Eq. (3.1). Provided that the set $\mathbf{z}_{t,i}$ is non-empty, the Dagenais estimator is consistent. Otherwise, we have perfect multicollinearity and β is not identified. This is a difference between data aggregation and general missing data problems.

There is one obvious problem with this estimator — the imputed $\widehat{x}_{t,1}, ..., \widehat{x}_{t,n}$ cannot sum up to $\overline{x}_t$. Therefore, the information content in the aggregated data is not fully explored.

The third is the minimum mean squared error (MSE) two-step estimator proposed by Hsiao (1979). The estimator is similar to the Dagenais estimator except that the imputed value is given by

$$\widehat{x}_{t,i} = \mathbf{z}_{t,i}\widehat{\alpha} + \mathbf{w}_{t,i}\widehat{\gamma} + \frac{1}{n}\left(\overline{x}_t - \sum_{i=1}^{n} \mathbf{z}_{t,i}\widehat{\alpha} + \mathbf{w}_{t,i}\widehat{\gamma}\right).$$

Essentially, we spread the imputation discrepancy $\overline{x}_t - \sum_{i=1}^{n} \mathbf{z}_{t,i}\widehat{\alpha} + \mathbf{w}_{t,i}\widehat{\gamma}$ evenly across the fitted value $\mathbf{z}_{t,i}\widehat{\alpha} + \mathbf{w}_{t,i}\widehat{\gamma}$. By construction, the aggregation constraint is always satisfied. The rationale of the imputation can be seen in Proposition 3.1. The imputed value is the conditional mean of $x_{t,i}\,|\overline{x}_t$, hence the name "minimum MSE". Furthermore the covariance structure of $\mathbf{x}_{t,-n}\,|\overline{x}_t$ implies the negative correlation of imputation errors. Therefore, Hsiao (1979) proposed using GLS in the second step regression. When $\sigma_{uv} = 0$, the covariance matrix of the disturbances is block diagonal, with the covariance within a group (block) given by $\beta^2\sigma_v^2\left(\mathbf{I}_n - \frac{1}{n}\iota_n\iota_n'\right) + \sigma_u^2\mathbf{I}_n$.

Although it seems that the minimum MSE estimator makes the best use of the information and should outperform the other two estimators, in fact that none of the three LS estimators dominates the others. First, if $\sigma_{uv} \neq 0$, the minimum MSE estimator is inconsistent due to endogeneity, while the Dagenais estimator is still consistent. Second, if imputation is of poor quality — σ_v^2 is large, it is possible that Dagenais estimator is less efficient than the all-aggregated-data estimator. More details are provided in the appendix.

Lastly, since both the Dagenais estimator and the minimum MSE estimator replace the true $x_{t,i}$ with the imputed value $\widehat{x}_{t,i}$ in Eq. (3.1), the conventional OLS standard error underestimates the true variability of the estimator. One solution is to analytically derive a modified standard error by accounting for all uncertainties, another strategy is to use multiple imputation. In the latter method, we sample $\left(\widehat{\alpha}^*, \widehat{\gamma}^*, \widehat{\sigma}_v^{2*}\right)$ from the distribution of $\left(\widehat{\alpha}, \widehat{\gamma}, \widehat{\sigma}_v^2\right)$, and then

generate the noise term $v_{t,i}^*$ from $N\left(0, \widehat{\sigma}_v^{2*}\right)$. Therefore, one set of simulated "complete data" for the Dagenais estimator is constructed as

$$\widehat{x}_{t,i}^* = \mathbf{z}_{t,i}\widehat{\alpha}^* + \mathbf{w}_{t,i}\widehat{\gamma}^* + v_{t,i}^*.$$

Similarly, the simulated "complete data" for the minimum MSE estimator is

$$\widehat{x}_{t,i}^* = \mathbf{z}_{t,i}\widehat{\alpha}^* + \mathbf{w}_{t,i}\widehat{\gamma}^* + v_{t,i}^* + \frac{1}{n}\left(\overline{x}_t - \sum_{i=1}^{n} \mathbf{z}_{t,i}\widehat{\alpha}^* + \mathbf{w}_{t,i}\widehat{\gamma}^* + v_{t,i}^*\right).$$

Repeat the process several times, we obtain several copies of the "complete data". For each copy, we estimate Eq. (3.1) by OLS. The final point estimator is the average of repeated estimates, with the total variance equal to the variance of repeated estimates (the between variability), plus the average of the estimated variances (the within variability).

3.6 Simulation studies

In this section, we use simulated data to evaluate the performance of various estimators listed in previous sections. For the ACD model with likelihood separability, we compare the analytic ML estimator to three LS estimators, focusing on their relative efficiency. For the model without separability, we compare the performance of the numerical ML and the Gibbs sampler, with the focus on the estimator stability.

For the case with separability, the simulated data experiment is specified as follows:

$n = 12$, $T = 300$,

$y_{t,i} = (x_{t,i}, \mathbf{w}_{t,i}) \cdot (1, 2, 3, 4)' + u_{t,i}$,

$x_{t,i} = (z_{t,i}, \mathbf{w}_{t,i}) \cdot \left(\frac{1}{2}, 1, 1, 1\right)' + v_{t,i}$,

$$\begin{pmatrix} u_{t,i} \\ v_{t,i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.1 \end{pmatrix}\right].$$

$z_{t,i}$ and three components of $\mathbf{w}_{t,i}$ are generated from i.i.d. $N\left(0, \frac{1}{4}\right)$.

For each set of simulated data, we obtain the three LS estimators (i.e. the all-aggregated-data, Dagenais, and minimum MSE estimators) as well as the analytic ML estimator. Then we repeat the data generating process 500 times, hence 500 copies of estimators.

Summary statistics are reported in Table 3.1. Each estimation approach takes three columns. The first column reports the estimator corresponding to the first simulated data set. The second column shows the average point estimators in 500 repetitions. The third column lists the standard deviation of the 500 repetitions, which can be viewed as the Monte Carlo standard error of the point estimator. If divided by $\sqrt{500}$, it indicates the numerical standard error (NSE) of the average estimator.

In the current setting, $\sigma_{uv} \neq 0$, the minimum MSE estimator is inconsistent (see Section 3.5). The point estimator $\widehat{\beta}$ averages 1.115 with the NSE 0.0038, significantly different from the true value of 1. Similarly, $\widehat{\delta}$ is also biased due to the endogeneity of $x_{t,i}$.

The simulation results also confirm that both the Dagenais estimator and ML estimator are consistent. $\widehat{\beta}$ using the Dagenais imputation averages 0.998 with the NSE 0.0036, and ML has an average value of 0.996 and NSE 0.0031. Both are close to the true value. However, the Dagenais estimator neglects the aggregation constraint and information usage is inadequate, so we observe that the Monte Carlo standard error of Dagenais estimator is 0.081, which is larger than that of the ML estimator which is 0.069.

The presence of the correlation between disturbances across equations biases the minimum MSE estimator and all-aggregated-data OLS estimator. However, both are consistent when $\sigma_{uv} = 0$. Results when σ_{uv} is changed from 0.1 to 0 are shown in Table 3.2. On average, $\widehat{\beta}$ for the all-aggregated-data, Dagenais, and minimum MSE estimators are 1.001, 0.997, 0.997 respectively. However, the standard errors are 0.105, 0.084, 0.076 respectively. The minimum MSE estimator incorporates the information content of both $z_{t,i}\widehat{\alpha}$ and $\overline{x}_t$, and therefore outperforms the other two. Also note that the likelihood is not separable when $\sigma_{uv} = 0$, if we still use analytic ML we have to estimate σ_{uv} as well, which is a source of efficiency loss.

	OLS			Dagenais			Minimum MSE			ML		
	1st	mean	std	1st	mean	std	1st	mean	std	1st	mean	std
β	1.576	1.614	0.099	1.008	0.998	0.081	1.133	1.115	0.085	0.983	0.996	0.069
δ_1	1.425	1.386	0.123	1.992	1.999	0.091	1.866	1.882	0.095	2.019	2.003	0.077
δ_2	2.312	2.391	0.124	2.968	3.004	0.091	2.842	2.887	0.096	2.970	3.007	0.076
δ_3	3.421	3.392	0.130	3.921	4.004	0.091	3.786	3.887	0.096	3.985	4.007	0.078
σ_u^2	0.412	0.434	0.036	0.412	0.434	0.036	0.412	0.434	0.036	0.497	0.500	0.026
α_1				0.483	0.502	0.036	0.483	0.502	0.036	0.495	0.502	0.028
γ_1				1.011	1.003	0.036	1.011	1.003	0.036	1.010	1.002	0.027
γ_2				0.993	1.000	0.037	0.993	1.000	0.037	1.017	0.999	0.027
γ_3				1.075	0.998	0.038	1.075	0.998	0.038	1.037	0.997	0.027
σ_v^2				0.105	0.099	0.008	0.105	0.099	0.008	0.108	0.100	0.007
σ_{uv}				0.153	0.135	0.026	0.124	0.110	0.024	0.115	0.101	0.010

The results are based on 500 simulations. Each estimation approach takes three columns. The first column reports the estimator for the first simulated data set. The second and third column show the average and standard deviation of the 500 repetitions. For the aggregated data OLS estimator, only Eq. (3.1) is estimated. For the Dagenais and minimum MSE estimator, Eq. (3.2) is regressed identically, so the numbers are the same. The estimated σ_{uv} is obtained from the identity: $Var\left(\beta v_{t,i} + u_{t,i}\right) = \beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}$, where $Var\left(\beta v_{t,i} + u_{t,i}\right)$ is estimated from the regression of $\{y_{t,i}\}$ on $\{z_{t,i}, \mathbf{w}_{t,i}\}$, σ_v^2 is estimated from regressing $\{\overline{x}_t\}$ on $\{\overline{z}_t, \overline{\mathbf{w}}_t\}$, and σ_u^2 from regressing $\{\overline{y}_t\}$ on $\{\overline{x}_t, \overline{\mathbf{w}}_t\}$. In the current setting with $\sigma_{uv} = 0.1$, only the Dagenais and ML estimators are consistent.

Table 3.1 Monte Carlo comparsion of LS and ML estimators, $\sigma_{uv} \neq 0$

	OLS			Dagenais			Minimum MSE			ML		
	1st	mean	std	1st	mean	std	1st	mean	std	1st	mean	std
β	1.034	1.001	0.105	1.001	0.997	0.084	1.015	0.997	0.076	0.985	0.997	0.081
δ_1	1.972	2.001	0.130	1.994	2.000	0.093	1.980	2.000	0.086	2.012	2.002	0.090
δ_2	2.853	3.003	0.128	2.957	3.005	0.093	2.943	3.005	0.086	2.960	3.006	0.090
δ_3	4.030	4.002	0.137	3.931	4.006	0.093	3.917	4.006	0.087	3.975	4.006	0.091
σ_u^2	0.468	0.495	0.041	0.468	0.495	0.041	0.468	0.495	0.041	0.486	0.499	0.025
α_1				0.490	0.503	0.037	0.490	0.503	0.037	0.498	0.503	0.035
γ_1				1.011	1.003	0.037	1.011	1.003	0.037	1.009	1.002	0.034
γ_2				1.017	0.999	0.037	1.017	0.999	0.037	1.031	0.998	0.034
γ_3				1.068	0.997	0.038	1.068	0.997	0.038	1.041	0.997	0.035
σ_v^2				0.103	0.099	0.008	0.103	0.099	0.008	0.104	0.099	0.008
σ_{uv}				0.025	0.004	0.022	0.023	0.004	0.021	0.017	0.002	0.014

The results are based on 500 simulations. Each estimation approach takes three columns. The first column reports the estimator for the first simulated data set. The second and third column show the average and standard deviation of the 500 repetitions. For the aggregated data OLS estimator, only Eq. (3.1) is estimated. For the Dagenais and minimum MSE estimator, Eq. (3.2) is regressed identically, so the numbers are the same. The estimated σ_{uv} is obtained from the identity: $Var\left(\beta v_{t,i}+u_{t,i}\right)=\beta^2\sigma_v^2+\sigma_u^2+2\beta\sigma_{uv}$, where $Var\left(\beta v_{t,i}+u_{t,i}\right)$ is estimated from the regression of $\{y_{t,i}\}$ on $\{z_{t,i},\mathbf{w}_{t,i}\}$, σ_v^2 is estimated from regressing $\{\overline{x}_t\}$ on $\{\overline{z}_t,\overline{\mathbf{w}}_t\}$, and σ_u^2 from regressing $\{\overline{y}_t\}$ on $\{\overline{x}_t,\overline{\mathbf{w}}_t\}$. In the current setting with $\sigma_{uv}=0$, all estimators are consistent.

Table 3.2 Monte Carlo comparsion of LS and ML estimators, $\sigma_{uv}=0$

When the likelihood is not separable, we compare the performance of the Newton-type numerical ML and Bayesian estimator using the Gibbs sampler. We consider a model without separability by adding one covariate in $\mathbf{z}_{t,i}$. $x_{t,i}=(\mathbf{z}_{t,i},\mathbf{w}_{t,i})\cdot(1,1,1,1,1)'+v_{t,i}$. Other settings remain the same.

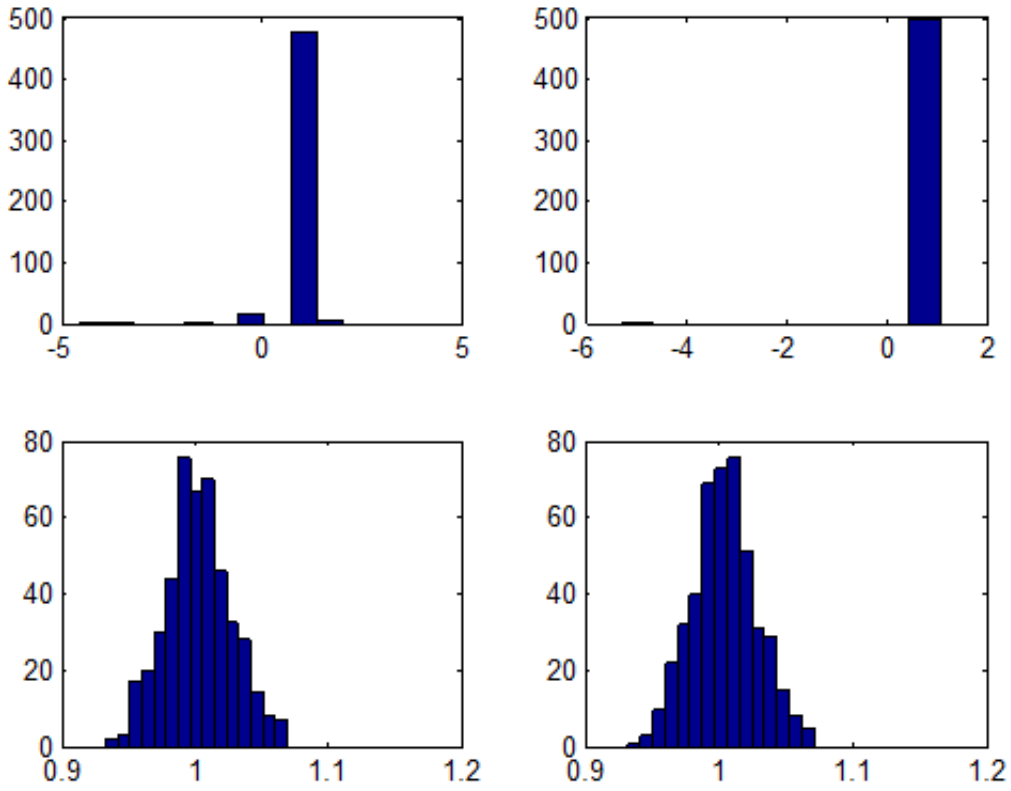
Though the traditional and Bayesian inference differ fundamentally on the parameter uncertainty, both of the them fully use the sampling information. If the priors are rather diffuse, Bayesian inference should also rely on the full-information likelihood function, and thus in large samples the posterior mean (or mode) should be close to the ML estimator and the posterior standard deviation close to the ML standard error. Here the major concern is to determine which numerical procedure can lead to ideal results in terms of speed and stability.

We specify the prior as $\psi\sim N\left(\mathbf{0},100\cdot\mathbf{I}_9\right)$, $\mathbf{\Sigma}^{-1}\sim Wishart\left(\mathbf{I}_2,1\right)$, which contains little information compared with the likelihood. The Gibbs sampler is run for 20000 draws with the

		Numeric ML			Bayesian		
	True	1st	mean	std	1st	mean	std
β	1	1.023	1.008	0.060	1.025	1.004	0.024
δ_1	2	1.964	1.993	0.075	1.963	1.998	0.042
δ_2	3	2.997	2.995	0.081	2.997	2.997	0.045
δ_3	4	3.973	3.991	0.077	3.971	3.995	0.041
σ_u^2	0.5	0.493	0.506	0.067	0.495	0.500	0.021
α_1	1	1.013	0.997	0.041	1.013	0.999	0.026
α_2	1	0.999	0.995	0.040	0.998	0.997	0.025
γ_1	1	0.978	1.000	0.030	0.977	0.999	0.025
γ_2	1	0.987	0.999	0.038	0.985	1.000	0.026
γ_3	1	0.968	1.001	0.039	0.967	1.001	0.028
σ_v^2	0.1	0.095	0.101	0.017	0.099	0.104	0.007
σ_{uv}	0.1	0.099	0.096	0.039	0.097	0.098	0.008

The results are based on 500 simulated data sets. The summary statistics are calculated with the apparent outliers ($\widehat{\beta} < 0$ or $\widehat{\beta} > 2$) removed. Each estimation approach takes three columns. The first column reports the estimator corresponding to the first simulated data set. The second column shows the average of the 500 repetitions. The third column lists the standard deviation of the 500 repetitions.

Table 3.3 Monte Carlo comparison of ML and Bayesian estimators



Histogram of $\widehat{\beta}$ in 500 repetitions. The upper left panel shows the ML estimation, and the upper right panel is the Bayesian estimation. The bottom left histogram truncate $\widehat{\beta}$ to (0.93, 1.07) for ML, and bottom right for Bayesian estimator.

Figure 3.1 A comparison of ML and Bayesian estimators to the ACD model

first half of draws burned in. The convergence and mixing diagnostics reveal that the chain has converged. We treat the posterior mean as the Bayesian point estimator.

We first generate a simulated dataset and set the initial values by adding a $N(0,1)$ disturbance on each true parameter value. If $\mathbf{\Sigma}$ is not positive definite, another disturbance draw is taken. Then the generated data and initial values are applied to both ML and the Gibbs sampler. Finally, the process is repeated for 500 times.

As is known, numerical ML can be sensitive to the initial values. In the 500 repetitions of simulated datasets, the numerical ML crashes 14 times and yields another 8 estimates departing far from the true values. Since the estimator standard deviation is no more than 0.1 and the true value of β is one, we define abnormality to be $\widehat{\beta} \leq 0$ or $\widehat{\beta} \geq 2$. To visualize the departing pattern of abnormal estimators, Figure 3.1 presents the histogram of $\widehat{\beta}$ in the 500 repetitions. In the case of crash, $\widehat{\beta} = 0$ is assigned for histogram presentation purpose. Compared with the numerical ML, the Gibbs sampler is more stable. It does not crash, and only yields negative $\widehat{\beta}$ twice. The abnormal estimators in the Gibbs sampler are close to each other. It is likely the chain gets stuck in a local high density region and cannot transverse to the region where the true parameters are located.

With the abnormal estimators removed, the summary statistics are presented in Tables 3.3. The average of the ML and Bayesian estimates are reasonably close to each other. Estimates of $\widehat{\beta}$ average 1.008 for ML and 1.004 for Bayesian. But the standard deviation of the ML estimates is 0.060, larger than that of the Bayesian 0.024. Though the role of the prior distribution and finite draws of the Gibbs sampler may partially explain the smaller variance of the estimator, we do not believe it is the main reason. The numerical issues should be taken into account. With obvious outliers removed, all of the Bayesian estimates lie in (0.93, 1.07) which is about plus/minus 3 standard deviations of the average point estimator. However, we observe several ML estimates with values such as 0.88, 1.22, 1.67. It is not clear it is caused by non-convergence of the optimizer or just caused by the sampling variation. Those values certainly exert a non-negligible impact on the calculation of the sample mean and standard deviation of the estimators. If we truncate the ML $\widehat{\beta}$ in the region (0.93, 1.07) as well, the mean is 1.003 with standard deviation 0.025, which is closer to the inference under the Bayesian scheme.

Although the numerical ML is not as stable as the Gibbs sampler, it does run faster. The average ML estimation time for one set of simulated data is about 1.04 seconds, while the 20000 draws with the Gibbs sampler takes an average of 37.9 seconds on an ordinary desktop computer (2.5GHz CPU / 3GB RAM / MATLAB 2009b). Nevertheless, the computation costs for both methods are affordable.

3.7 Extensions

In an empirical context, the problems we have encountered might be more complicated than the baseline ACD model. In this section, we outline several extensions to the model and ways to handle them.

3.7.1 Aggregation of several variables

If more than one covariate is aggregated, the model can be extended as

$$\begin{aligned} y_{t,i} &= x_{1t,i}\beta_1 + ...x_{kt,i}\beta_k + \mathbf{w}_{t,i}\delta + v_{0t,i} \\ x_{1t,i} &= \mathbf{z}_{t,i}\alpha_1 + \mathbf{w}_{t,i}\gamma_1 + v_{1t,i} \\ &... \\ x_{kt,i} &= \mathbf{z}_{t,i}\alpha_k + \mathbf{w}_{t,i}\gamma_k + v_{kt,i} \end{aligned}$$

with $\overline{x}_{1t} = \sum_{i=1}^{n} x_{1t,i}$, ..., $\overline{x}_{kt} = \sum_{i=1}^{n} x_{kt,i}$, $(v_{0t,i}, v_{1t,i}, ..., v_{kt,i}) \sim N(\mathbf{0}, \mathbf{\Omega})$.

For ML, we maximize the joint density of observable variables:

$$\ln L(\theta) = \sum_{t=1}^{T} \ln f(y_{t,1}, ..., y_{t,n}, \overline{x}_{1t}, ..., \overline{x}_{kt}).$$

The model can always be estimated by numerical ML. However, for analytic solution we require $\mathbf{z}_{t,i}$ contains exactly k variables. The likelihood can be factorized as

$$f(\mathbf{y}_t, \overline{x}_{1t}, ..., \overline{x}_{kt}) = f(\mathbf{y}_t) \cdot f(\overline{x}_{1t} | \mathbf{y}_t) \cdot ... \cdot f\left(\overline{x}_{kt} \left| \mathbf{y}_t, \overline{x}_{1t}, ..., \overline{x}_{(k-1)t}\right.\right).$$

The analytic ML estimator is obtained from $k+1$ auxiliary regressions:

Regress $y_{t,i}$ on $\mathbf{z}_{t,i}$, $\mathbf{w}_{t,i}$.

Regress $\overline{x}_{1t}$ on $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$.

Regress $\overline{x}_{2t}$ on $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$, $\overline{x}_{1t}$.

... ...

Regress $\overline{x}_{kt}$ on $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$, $\overline{x}_{1t}$,..., $\overline{x}_{(k-1)t}$.

Lastly, by solving an equation system we can recover the ML estimator of $\beta_1, ..., \beta_k$, δ, $\alpha_1, ..., \alpha_k$, $\gamma_1, ..., \gamma_k$, $\mathbf{\Omega}$. See appendix for a derivation.

Suppose $\mathbf{z}_{t,i}$ contains J variables ($J < k$), and some of the covariance terms in $\mathbf{\Sigma}$ are restricted to zero, we still have analytic solution. For example, if we believe $x_{1t,i}$ is uncorrelated with $v_{0t,i}$, then the first row and first column of $\mathbf{\Sigma}$, except for the diagonal element, are restricted to zero. The reparameterized estimator is obtained from the same regressions as above, but $\mathbf{z}_{t,i}$ and $\overline{\mathbf{z}}_t$ have reduced dimensions.

In words, separability requires that one instrument corresponds to one endogenous variable.

3.7.2 Unbalanced aggregation

In some applications, group sizes are not equal so that n needs to be written as n_t.

Compared with the ML solutions outlined in Section 3, the separability conditions do not change. The expression of $f(\mathbf{y}_t)$ remains the same, so does the first regression of $y_{t,i}$ on $\mathbf{z}_{t,i}$, $\mathbf{w}_{t,i}$. However, the variance of $\overline{x}_t | \mathbf{y}_t$ changes:

$$f(\overline{x}_t | \mathbf{y}_t) = \phi(\overline{x}_t; \overline{z}_t \cdot D + \overline{\mathbf{w}}_t \cdot \mathbf{E} + \overline{y}_t \cdot F, n_t G),$$

where $n_t G = n_t \sigma_v^2 - n_t \left(\beta \sigma_v^2 + \sigma_{uv}\right)^2 \left(\beta^2 \sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}$. Other components remain the same as before. It implies that

$$\max_{D,\mathbf{E},F,G} \sum_{t=1}^{T} \ln f(\overline{x}_t | \mathbf{y}_t)$$

can be obtained by weighed least squares of $\overline{x}_t$ on $\overline{z}_t, \overline{\mathbf{w}}_t, \overline{y}_t$ with the weights proportional to n_t.

Procedures of the Bayesian simulator are largely unchanged in the unbalanced aggregation. Full posterior conditionals of ψ and $\mathbf{\Sigma}^{-1}$ remain the same, and we use n_t instead of n when groupwise taking draws of $\{x_{t,i}\}$.

3.7.3 Partial aggregation

In applied problems, data aggregation may take on another degree of complexity. For instance, in Example 2, suppose we do find county-level pet population for some states, but not for the rest. How do we make the best use of the incomplete county-level data instead of regressing merely with aggregated state-level data?

In general, the partial aggregation problem is raised as follows: suppose group t has n_t members, among which the first k_t are observable and the rest are missing. In addition, the aggregated value $\overline{x}_t = \sum_{i=1}^{n} x_{t,i}$ is known. The data are generated according to Eq. (3.1) and (3.2).

To address this problem, group t can be divided into $k_t + 1$ smaller groups. The first k_t groups are a singleton with known $x_{t,i}$, and the last group contains $n_t - k_t$ members, whose latent values sum up to $\overline{x}_t - \sum_{i=1}^{k_t} x_{t,i}$. Then the problem is equivalent to the unbalanced aggregation introduced in the previous subsection, and both ML and Bayesian estimators can be implemented.

3.8 Conclusion

Hsiao's model offers a simple framework for addressing data aggregation problems. This paper explores several estimation strategies for this model, showing that the solutions do not always require numerical tools proposed by Palm and Nijman (1982).

The first is a full-information ML estimation. We find that the likelihood function has a separability property in two useful special cases. As long as one instrument corresponds to one endogenous variable, the likelihood can be maximized analytically, with the ML estimator obtained by two linear regressions. For models without the separable likelihood, numerical procedures can also be used, but initial values must be carefully chosen.

The second is the Bayesian simulator implemented by the Gibbs sampler, the advantage of which is two-fold. First, it is more stable. Our Monte Carlo study shows that the Bayesian estimator is less affected by the initial values. Second, it is more flexible. It places no re-

strictions on the covariates in the imputation regression, and the sampling procedure of latent disaggregated covariates can be easily inserted into researchers' models.

The third is a class of LS estimators. The Dagenais two-step estimator is primarily used for imputing missing data, but it is suitable for aggregated covariate data as well. The minimum MSE estimator is based on the regression imputation, but also uses the aggregation constraint. In the absence of correlation of disturbances among equations, the latter makes better use of information and yields a more precise imputation. Otherwise, the latter is inconsistent, but the former is still consistent. On top of that, the all-aggregated-data OLS method offers the simplest way to estimate the model, and is useful when the imputation is of poor quality. Though conceptually LS is easier to implement than ML and the Gibbs sampler, it is not as efficient in general and thus not recommended unless we cast doubt on the normality assumption.

CHAPTER 4. SAMPLING VARIATION, MONOTONE INSTRUMENTAL VARIABLES AND THE BOOTSTRAP BIAS CORRECTION

4.1 Introduction

Proposed by Manski and Pepper (2000), Monotone instrumental variables (MIV) is a powerful tool for treatment response identification. The MIV assumption weakens the traditional instrumental variable assumption by a weak inequality of mean response across sub-populations. As a result, the MIV sharp lower bound invariably involves a supremum operator and the upper bound contains an infimum operator.

However, when sampling variation is taken into account, the bounds themselves assume randomness since the population moments or probabilities are replaced by their analogues. Though the analogue estimates are still consistent, finite sample bias is a serious concern. As is noted by Manski and Pepper (2009, p.211), "the sup and inf operations ... significantly complicate the bounds under other MIV assumptions, rendering it difficult to analyze the sampling behavior of analogue estimates."[1] The major statistical problem is that the analogue estimate of the lower bound is biased upwards and upper bound biased downwards, resulting in the estimates narrower than the true bounds.

To address this concern, two major lines of research are present in the literature to the best of our knowledge. One is direct adjustment. Chernozhukov et al. (2009) develop an inference method on intersection bounds with a continuum of inequalities. Their estimator maximizes or minimizes the precision-corrected curve defined by the analogue estimates plus a critical value multiplied by pointwise standard errors. Another solution is bootstrap adustment. Kreider

[1]The bounds under the monontone treatment selection assumption have simple forms, but under other MIV assumptions the supremum and infimum operators will appear in the bounds.

and Pepper (2007) propose a heuristic bootstrap bias correction and applied this approach to their employment gap identification problems. Though Monte Carlo experiments in Manski and Pepper (2009) provide evidence on the effectiveness of bias reduction, theoretical foundation has not been established to justify the bootstrap correction. In addition, the simulation results of Manski and Pepper (2009) show that in some cases moderate biases remain after the correction.

The goal of this paper is to justify the bootstrap bias correction. Traditionally, the improvement of the corrected estimator is in the sense of asymptotic refinement. That is, we expect the bootstrap corrected estimator has a bias going to zero at a faster rate than the uncorrected estimator. However, there are difficulties applying asymptotic expansion techniques to our problem, since the bounds under the MIV assumption are not differentiable. In this paper, we take an innovative, and perhaps more direct, approach to study bootstrap bias reduction. We rely on asymptotic normality of the estimators to derive our results. Given normally distributed variates, we bound the magnitude of the upward bias induced by the $\max(\cdot)$ operator and show how the one-level bootstrap reduces this upward bias but cannot eliminate it. In some circumstances, one-level bootstrap may over-correct the bias. Then under an assumption that the bias function can be approximated by a polynomial, we show the mechanism of the multi-level bootstrap bias correction, which successively lowers the order of the polynomial towards unbiasness. Lastly, to make multi-level bootstrap computationally feasible, we propose a simultaneous bootstrap procedure which conducts many levels of bootstraps at affordable computational costs.

For convenience, we discretize every random variable so that we can use a categorical distribution of several dimensions to characterize their joint distribution, which makes easier the statistical properties of the analogue MIV bounds. For this problem, discretization is not unreasonable. First, the treatment variable is discrete, usually binary, in most applications. Second, MIV identification requires that the response variable is bounded below and above. Otherwise the MIV has no identification power unless it is used together with monotone treatment selection defined in Manski (1997). Finite-valued discrete distribution by nature has a lower and upper bound. Third, to compute the analogue estimates for each subpopulation classified by MIV, we usually group the values of the MIV so as to ensure sufficient sample size. Therefore,

we model treatments, responses and MIVs as finite-valued discrete random variables.

4.2 The mathematical structure of MIV bounds

Manski and Pepper (2000, 2009) use MIV to help bound counterfactual outcomes, while Kreider and Pepper (2007) consider MIV identification in a partial misreporting problem. Though the derived MIV bounds look different, they share the same mathematical structure, so the same bias correction procedure can be applied to both problems. In this section, we summarize their common structure.

The counterfactual outcomes identification problem can be raised as follows. Let $D \in \{d_1, ..., d_{n_D}\}$ be a treatment variable. The n_D varieties of treatments generate n_D types of latent responses, denoted as $Y_t \in \{y_1, ..., y_{n_Y}\}$, $t = 1, ..., n_D$. Since a person cannot receive more than one treatment simultaneously, the only observable outcome is $Y = \sum_{t=1}^{n_D} Y_t \cdot I(D = d_t)$, where $I(\cdot)$ is an indicator function. Let $Z \in \{z_1, ...z_{n_Z}\}$ be a MIV such that for any two realizations $z_i \leq z_j$,

$$E(Y_t | Z = z_i) \leq E(Y_t | Z = z_j), \forall t = 1, ..., n_D.$$

Without loss of generality, discrete values of Y_t and Z are sorted in an increasing order: $y_1 \leq y_2 ... \leq y_{n_Y}$, $z_1 \leq z_2 ... \leq z_{n_Z}$.

Consider $E(Y_t | Z = z_j)$ for some $t = 1, ..., n_D$, $j = 1, ..., n_Z$. It is bounded below by $\sup_{1 \leq i \leq j} E(Y_t | Z = z_i)$ and above by $\inf_{j \leq i \leq n_Z} E(Y_t | Z = z_i)$. Since the MIV is discretized, we can replace $\sup(\cdot)$ by $\max(\cdot)$, and $\inf(\cdot)$ by $\min(\cdot)$. Furthermore, $E(Y_t | Z = z_i)$ can be dissembled into an observable part $E(Y | Z = z_i, D = d_t)$ and an unobservable part $E(Y_t | Z = z_i, D \neq d_t)$. The latter needs to be replaced by the worse-case lower bound y_1 and upper bound y_{n_Y}, which yields the sharp bounds under the MIV assumption alone:

$$\begin{aligned} &\max_{1 \leq i \leq j} E(Y | Z = z_i, D = d_t) \cdot P(D = d_t | Z = z_i) + y_1 \cdot P(D \neq d_t | Z = z_i) \\ &\leq E(Y_t | Z = z_j) \leq \\ &\min_{j \leq i \leq n_Z} E(Y | Z = z_i, D = d_t) \cdot P(D = d_t | Z = z_i) + y_{n_Y} \cdot P(D \neq d_t | Z = z_i). \end{aligned} \tag{4.1}$$

To make notation compact, let us define

$p_{ikm} \equiv P(Z = z_i, Y = y_k, D = d_m)$,

$i = 1, ..., n_Z$, $k = 1, ..., n_Y$, $m = 1, ..., n_D$,

$p_{i\cdot\cdot} \equiv \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} p_{ikm}$,

$\mathbf{p} \equiv vec\left(\{p_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}\right)$,

$\mathbf{p}_i \equiv vec\left(\{p_{ikm}\}_{k=1,m=1}^{n_Y,n_D}\right)$.

Here $vec(\cdot)$ is an operator that vectorizes a multi-dimension array into a long column vector. For instance, $vec\left(\{p_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}\right)$ turns a $n_Z \times n_Y \times n_D$ array to a $n_Z n_Y n_D \times 1$ vector. Also assume $p_{i\cdot\cdot} > 0$, $\forall i = 1, ..., n_Z$. Then we can rewrite Eq. (4.1) as

$$\max_{1 \le i \le j} f_L(\mathbf{p}_i) \le E(Y_t | Z = z_j) \le \min_{j \le i \le n_Z} f_U(\mathbf{p}_i), \tag{4.2}$$

where

$$f_L(\mathbf{p}_i) = \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} \frac{p_{ikm}}{p_{i\cdot\cdot}} [y_k \cdot I(m = t) + y_1 \cdot I(m \neq t)],$$

$$f_U(\mathbf{p}_i) = \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} \frac{p_{ikm}}{p_{i\cdot\cdot}} [y_k \cdot I(m = t) + y_1 \cdot I(m \neq t)].$$

The misreporting identification problem in Kreider and Pepper (2007) uses respondents' self-reported health information to bound the effects of (true) disability on employment. Let $L \in \{0, 1\}$ be observed employment status, $X \in \{0, 1\}$ and $W \in \{0, 1\}$ be the reported and true disability status respectively, and $Z \in \{z_1, ...z_{n_Z}\}, z_1 \le z_2 ... \le z_{n_Z}$ be a MIV (namely negative age in their paper) such that

$$P(L = 1 | W, Z = z_i) \le P(L = 1 | W, Z = z_j), \text{ if } i \le j.$$

Respondents are classified into two groups, namely the verified ($Y = 1$) and the unverified ($Y = 0$), on the basis of researchers' prior information on their accurate reporting rate. Taking this verification rate as given, Kreider and Pepper (2007) derive sharp bounds of $P(L = 1 | W = 1)$. For simplicity, we consider an extreme case that the verified group has a 100% truth-telling rate, while the unverified has an accuracy rate $\ge 0\%$ (i.e., no information). For each $j = 1, ..., n_Z$, we have

$$\max_{1\leq i\leq j} \frac{P\left(L=1,X=1,Y=1\,|Z=z_i\right)}{P\left(X=1,Y=1\,|Z=z_i\right)+P\left(L=0,Y=0\,|Z=z_i\right)} \tag{4.3}$$
$$\leq P\left(L=1\,|W=1,Z=z_j\right)\leq$$
$$\min_{j\leq i\leq n_Z} \frac{P\left(L=1,X=1,Y=1\,|Z=z_i\right)+P\left(L=1,Y=0\,|Z=z_i\right)}{P\left(X=1,Y=1\,|Z=z_i\right)+P\left(L=1,Y=0\,|Z=z_i\right)}$$

Readers are referred to Proposition 2, corollary 1 in Kreider and Pepper (2007, p.436) for the derivation. Note that when the accuracy rate is not as extreme as 100% and 0%, the bounds will be more cumbersome. However, what remains unchanged is that all the probabilities are conditional on $Z=z_i$. This feature makes the mathematical structure of the MIV bounds (see below) unchanged.

Define a set of symbols similar to what we defined in the previous problem.

$p_{ijkl}\equiv P\left(Z=z_i,L=j,X=k,Y=l\right),\ i=1,...,n_Z,\ j,k,l=0,1,$

$p_{i\cdots}\equiv\sum_{j=0}^{1}\sum_{k=0}^{1}\sum_{l=0}^{1}p_{ijkl},$

$\mathbf{p}\equiv vec\left(\{p_{ijkl}\}_{i=1,j=0,k=0,k=0}^{n_Z,1,1,1}\right),$

$\mathbf{p}_i\equiv vec\left(\{p_{ijkl}\}_{j=0,k=0,k=0}^{1,1,1}\right).$

Then Eq (4.3) can be written as

$$\max_{1\leq i\leq j} f_L\left(\mathbf{p}_i\right)\leq P\left(L=1\,|W=1,Z=z_j\right)\leq \min_{j\leq i\leq n_Z} f_U\left(\mathbf{p}_i\right), \tag{4.4}$$

where

$$f_L\left(\mathbf{p}_i\right)=\frac{p_{i111}}{p_{i111}+p_{i011}+p_{i010}+p_{i000}}$$
$$f_U\left(\mathbf{p}_i\right)=\frac{p_{i111}+p_{i110}+p_{i100}}{p_{i111}+p_{i011}+p_{i110}+p_{i100}}$$

Comparing Eq. (4.2) with Eq. (4.4), we see the MIV bounds of the two problems have some features in common:

First, the theoretical bounds are determined by $\mathbf{p}$, the parameter vector summarizing the joint probability of observable variates. In other words, the observable variates follows a categorical distribution of multiple dimensions, which is equivalent to a long single-dimension categorical distribution with parameters $\mathbf{p}$.

Second, the MIV bounds take the form of $\max_{1\leq i\leq j} f_L(\mathbf{p}_i)$ and $\min_{j\leq i\leq n_Z} f_U(\mathbf{p}_i)$, where $\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{n_Z}$ form a partition of $\mathbf{p}$ according to the possible values of the MIV.

Third, both $f_L(\mathbf{p}_i)$ and $f_U(\mathbf{p}_i)$ are homogeneous functions of degree zero. Eq. (4.1) and Eq. (4.3) involves probabilities conditional on $Z = z_i$, which is the ratio of the joint and the marginal probabilities. Since a constant cancels in the nominator and denominator, $f_L(\mathbf{p}_i)$ and $f_U(\mathbf{p}_i)$ in Eq. (4.2) and Eq. (4.4) always satisfy degree-zero homogeneity.

4.3 Sampling Variation

In applications, the probability vector $\mathbf{p}$ needs to be estimated from the data. Let $\{\mathbf{v}_s\}_{s=1}^n$ be i.i.d. draws from the categorical distribution with parameters $\mathbf{p}$. Conceptually, this means there are n persons taking the survey which asks for each respondent's realized choice of (Z, Y, D) or (Z, L, X, W). All possible choices of (Z, Y, D) define $n_Z n_Y n_D$ categories and that of (Z, L, X, W) define $8n_Z$ categories. So the length of the vector $\mathbf{v}_s$ is $n_Z n_Y n_D$ and $8n_Z$ respectively. The person s chooses a category, so the component in $\mathbf{v}_s$ corresponding to that realized category will be coded as 1 with other elements in $\mathbf{v}_s$ being 0.

By construction, the sample analogue of $\mathbf{p}$ can be expressed as

$$\widehat{\mathbf{p}} = \frac{1}{n}\sum_{s=1}^{n} \mathbf{v}_s.$$

Fact 4.1. *$\widehat{\mathbf{p}}$ is a consistent estimate of $\mathbf{p}$, and the asymptotic distribution is*

$$\sqrt{n}\left(\widehat{\mathbf{p}} - \mathbf{p}\right) \xrightarrow{d} N\left[\mathbf{0}, diag\left(\mathbf{p}\right) - \mathbf{p}\mathbf{p}'\right],$$

where $diag(\mathbf{p})$ refers to a diagonal matrix with the main diagonal being the vector $\mathbf{p}$.

Proof. See appendix. □

Suppose the length of $\mathbf{p}$ is r; then $diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'$ is a positive semidefinite matrix of reduced rank $r - 1$. The linear combination $\iota'\widehat{\mathbf{p}}$, where ι is a vector of ones, have the mean of one and variance of zero. Therefore, the analogue probability estimates always sum up to one. In addition, the elements of $\widehat{\mathbf{p}}$ are negatively correlated since they are subject to the aggregation constraint.

Fact 4.1 suggests that the large-sample approximating distribution of $\widehat{\mathbf{p}}$ is given by $N\left[\mathbf{p}, \frac{1}{n} diag\left(\mathbf{p}\right) - \frac{1}{n}\mathbf{p}\mathbf{p}'\right]$. Of course, it is understood that $\widehat{\mathbf{p}}$ is a bounded random vector since each component must fall in the unit interval.

Partition $\widehat{\mathbf{p}}$ into $\widehat{\mathbf{p}}_1, ..., \widehat{\mathbf{p}}_{n_Z}$ in the same way we partition $\mathbf{p}$ into $\mathbf{p}_1, ..., \mathbf{p}_{n_Z}$. Now we consider the asymptotic distribution of $f_L\left(\widehat{\mathbf{p}}_i\right)$, $f_U\left(\widehat{\mathbf{p}}_i\right)$, $i = 1, ..., n_Z$.

Proposition 4.2. *Let $f_L\left(\cdot\right)$ be a real differentiable function satisfying homogeneity of degree zero, that is, $f_L\left(c\mathbf{x}\right) = f_L\left(\mathbf{x}\right)$, $\forall c > 0$. Then $f_L\left(\widehat{\mathbf{p}}_1\right), ..., f_L\left(\widehat{\mathbf{p}}_{n_Z}\right)$ are asymptotically independent and for each $i = 1, ..., n_Z$,*

$$\sqrt{n}\left[f_L\left(\widehat{\mathbf{p}}_i\right) - f_L\left(\mathbf{p}_i\right)\right] \xrightarrow{d} N\left[\mathbf{0}, \mathbf{G}_i \cdot diag\left(\mathbf{p}_i\right) \cdot \mathbf{G}_i'\right],$$

where $\mathbf{G}_i$ is a row vector such that

$$\mathbf{G}_i = \frac{\partial f_L\left(\widehat{\mathbf{p}}_i\right)}{\partial \widehat{\mathbf{p}}_i'}\Big|_{\widehat{\mathbf{p}}_i = \mathbf{p}_i}.$$

Proof. See appendix. □

The asymptotic distribution of $f_U\left(\widehat{\mathbf{p}}_i\right)$ can be derived similarly with the subscript L replaced by U in Proposition 4.2.

The zero-degree homogeneity of $f_L\left(\cdot\right)$ plays an important role in Proposition 4.2 since Euler's Theorem implies that $\mathbf{G}_i\mathbf{p}_i = 0$, $i = 1, ..., n_Z$, resulting in both zero covariances and simplified variances of the normal variates. Theoretically, Proposition 4.2 provides a unified asymptotic distribution of $f_L\left(\cdot\right)$ for any identification problem with the MIV, as long as $f_L\left(\cdot\right)$ can be written as a differentiable function of the population probabilities conditional on the MIV. Proposition 4.2 will be also used to justify the assumptions of the bootstrap bias correction in the next section. Practically, Proposition 4.2 can be used to compute the asymptotically variance of $f_L\left(\widehat{\mathbf{p}}_i\right)$ if we are willing to calculate the cumbersome gradients. However, for a specific problem, there might be some better way to compute the finite-sample variance. For instance, once we recognize that the $f_L\left(\mathbf{p}_i\right)$ in Eq. (4.2) can be represented as a conditional expectation, the finite-sample variance of $f_L\left(\widehat{\mathbf{p}}_i\right)$ is readily given in the next proposition.

Proposition 4.3. *$f_L(\mathbf{p}_i)$ in Eq. (4.2) takes the following form:*

$$f_L(\mathbf{p}_i) = E(Q \,|\, Z = z_i),$$

where

$$Q = Y \cdot I(D = d_t) + y_1 \cdot I(D \neq d_t).$$

Conditional on the positive analogue $p_{i\cdot\cdot}$, the finite-sample variance of $f_L(\widehat{\mathbf{p}}_i)$ is given by

$$Var[f_L(\widehat{\mathbf{p}}_i)] = \left[\sum_{r=1}^{n} \frac{1}{r} \frac{\binom{n}{r}(p_{i\cdot\cdot})^r (1-p_{i\cdot\cdot})^{n-r}}{1-(1-p_{i\cdot\cdot})^n}\right] \cdot Var(Q \,|\, Z = z_i),$$

where

$$\begin{aligned} Var(Q \,|\, Z = z_i) &= E\left(Q^2 \,|\, Z = z_i\right) - [E(Q \,|\, Z = z_i)]^2 \\ &= \sum_{k=1}^{n_Y}\sum_{m=1}^{n_D} \frac{p_{ikm}}{p_{i\cdot\cdot}} q_{km}^2 - \left[\sum_{k=1}^{n_Y}\sum_{m=1}^{n_D} \frac{p_{ikm}}{p_{i\cdot\cdot}} q_{km}\right]^2, \end{aligned}$$

and

$$q_{km} = y_k \cdot I(d_m = d_t) + y_1 \cdot I(d_m \neq d_t).$$

Proof. See appendix. □

4.4 Estimating the MIV bounds

Proposition 4.2 indicates that the large-sample approximating distribution of $f_L(\widehat{\mathbf{p}}_i)$ is given by $N\left[f_L(\mathbf{p}_i), \frac{1}{n}\mathbf{G}_i \cdot diag(\mathbf{p}_i) \cdot \mathbf{G}_i'\right]$. To estimate the MIV bounds as in Eq. (4.2) and Eq. (4.4), we need to find an estimator for $\max_{1\leq i\leq j} f_L(\mathbf{p}_i)$. An naive choice is $\max_{1\leq i\leq j} f_L(\widehat{\mathbf{p}}_i)$. Though $f_L(\widehat{\mathbf{p}}_i)$ is an asymptotically unbiased estimator for $f_L(\mathbf{p}_i)$, $\max_{1\leq i\leq j} f_L(\widehat{\mathbf{p}}_i)$ is not an unbiased estimator for $\max_{1\leq i\leq j} f_L(\mathbf{p}_i)$ in the finite sample. It is biased upwards simply because $\max(\cdot)$ is convex and Jensen's inequality implies $E[\max_{1\leq i\leq j} f_L(\widehat{\mathbf{p}}_i)] > \max_{1\leq i\leq j} f_L(\mathbf{p}_i)$. Similarly, $\min_{j\leq i\leq n_Z} f_U(\widehat{\mathbf{p}}_i)$ has a downward bias if it is used to estimate $\min_{j\leq i\leq n_Z} f_U(\mathbf{p}_i)$. This is unfavorable from the perspective of decision making in that the estimated bounds are narrower than the true bounds. Kreider and Pepper (2007) propose a heuristic bootstrap bias correction. The Monte Carlo evidence in Manski and Pepper (2009) indicates the bias can be considerably reduced, but not eliminated after the correction. In this section, we will analyze the biases

of a series of estimators and provide a justification for the bootstrap correction. We will also suggest a feasible approach to conduct several levels of bootstraps simultaneously. We will focus on the bias correction of $\max_{1\leq i\leq j} f_L(\widehat{\mathbf{p}}_i)$, and the same principle can be applied to the case of $\min_{j\leq i\leq n_Z} f_U(\widehat{\mathbf{p}}_i)$ as well.

To make our notations compact, define

$\mu_i \equiv f_L(\mathbf{p}_i)$, $\sigma_i^2 \equiv \frac{1}{n}\mathbf{G}_i \cdot diag(\mathbf{p}_i) \cdot \mathbf{G}_i'$, $X_i \equiv f_L(\widehat{\mathbf{p}}_i)$, $i = 1, ..., j$.

$\boldsymbol{\mu} \equiv (\mu_1, ..., \mu_j)'$, $\boldsymbol{\sigma}^2 \equiv diag\left(\sigma_1^2, ..., \sigma_j^2\right)$, $\boldsymbol{X} \equiv (X_1, ..., X_j)'$.

Let $\mathbf{x}$ be a realization of $\boldsymbol{X}$. That is, the only one realized $\mathbf{x}$ is what we obtained from the data.

Essentially our task is to propose a good estimator for $\max(\boldsymbol{\mu})$ by observing $\mathbf{x}$. To that end, we need to make some assumptions.

Assumption 4.1. $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$.

Assumption 4.2. $\boldsymbol{\sigma}^2$ *is known.*

The rationale for the first assumption is Proposition 4.2, which suggests $X_1, ..., X_j$ are asymptotically independent normal variates. The second assumption is arguable. In practice, the variances of those variates are unknown, and we at best can provide a consistent estimator for the variances, say $\widehat{\boldsymbol{\sigma}}^2$, using Proposition 4.2 or Proposition 4.3. It is true that each σ_i^2 is positively related to the magnitude of the upward bias (which is most apparent if we assume the convex function is differentiable and examine the Taylor expansion). However, we do not know whether $E(\widehat{\sigma}_i^2)$ is larger or smaller than σ_i^2 in the finite sample, so at best we can argue that the upward bias derived with $\widehat{\sigma}_i^2$ will be close to the true upward bias determined by σ_i^2. In this sense, we view it as a working assumption.

4.4.1 Bias function and a conservative estimator

A naive estimator is the maximum of the sample.

$$T_1(\mathbf{x}) = \max(\mathbf{x}).$$

By Jensen's inequality, $E\left[T_1\left(\mathbf{X}\right)\right] > \max\left(\boldsymbol{\mu}\right)$. So the estimator is biased upwards. Define the first-level bias function $B_1 \colon \mathbb{R}^j \to \mathbb{R}$ such that

$$B_1\left(\boldsymbol{\mu}\right) = E\left[T_1\left(\mathbf{X}\right)\right] - \max\left(\boldsymbol{\mu}\right).$$

$B_1\left(\cdot\right)$ is a function of $\boldsymbol{\mu}$ since $\mathbf{X} \sim N\left(\boldsymbol{\mu},\boldsymbol{\sigma}^2\right)$. Of course, it is also a function of $\boldsymbol{\sigma}^2$, which is assumed to be known and therefore suppressed.

The first-level bias function has a useful property stated below.

Proposition 4.4 (Bounds of the bias function)**.** *$B_1\left(\cdot\right)$ is bounded by $0 < B_1\left(\boldsymbol{\mu}\right) \leq M$, $\forall$ $\boldsymbol{\mu} \in \mathbb{R}^j$, where*

$$M = E\left[\max\left(\mathbf{X}_0\right)\right],$$
$$\mathbf{X}_0 \sim N\left(\mathbf{0},\boldsymbol{\sigma}^2\right).$$

Proof. See appendix. □

Note that the upper bound M is computable, at least by simulation. For the special case of $j = 2$, Clark (1961), Cain (1994) provide an analytic result.

$$B_1\left(\boldsymbol{\mu}\right) = \omega\mu_1 + \left(1-\omega\right)\mu_2 + \sigma_0\phi\left(\frac{\mu_1-\mu_2}{\sigma_0}\right) - \max\left(\mu_1,\mu_2\right),$$
$$M = \sigma_0\phi\left(0\right),$$

where $\phi\left(\cdot\right)$, $\Phi\left(\cdot\right)$ is the standard normal p.d.f. and c.d.f. respectively, and

$$\omega = \Phi\left(\frac{\mu_1-\mu_2}{\sigma_0}\right),$$
$$\sigma_0 = \sqrt{\sigma_1^2+\sigma_2^2}.$$

For $j = 2$, we may plot a 3-D graph of $B_1\left(\cdot\right)$, with μ_1, μ_2 on the x, y axis and B_1 on the z axis (see Figure 4.1). It is a ridge-shaped function. Along the 45° line on the x, y plane, $B_1\left(\cdot\right)$ attains the same maximum value $\sigma_0\phi\left(0\right)$. Off the 45° line, $B_1\left(\cdot\right)$ gradually decreases towards zero.

Proposition 4.4 shows that the bias of the naive estimator $\max\left(\mathbf{X}\right)$ is bounded above, so we can propose a conservative estimator for $\max\left(\boldsymbol{\mu}\right)$:

$$T_c\left(\mathbf{x}\right) = \max\left(\mathbf{x}\right) - M.$$

By construction, T_c is biased downwards. We call it a conservative estimator because we can use the same principle to propose an upward biased estimator for $\min_{j\leq i\leq n_Z} f_U(\widehat{\mathbf{p}}_i)$, and then we will obtain bounds wider than the true bounds. For decision making, perhaps we would rather have too wide bounds than too narrow bounds. Also note that if we allow $\boldsymbol{\sigma}^2 \to \mathbf{0}$, M will also decrease to zero, so that T_c will converge to $\max(\boldsymbol{\mu})$. Therefore, if T_c is applied to the MIV bounds, it is still a consistent estimator. Furthermore, since T_1 is biased upwards and T_c is biased downwards, they themselves bound the unbiased estimator of the MIV bounds.

4.4.2 Bootstrap bias correction

Clearly, T_c over-corrects the bias. Is it possible to find an estimator "being just right"? Kreider and Pepper (2007) proposed a heuristically motivated bootstrap bias corrected estimator. This subsection aims to provide a rationale for this correction.

The idea of bootstrap bias correction is to use the bias function to correct the naive estimator. Define

$$T_2^*(\mathbf{x}) = T_1(\mathbf{x}) - B_1(\boldsymbol{\mu}),$$

$$T_2(\mathbf{x}) = T_1(\mathbf{x}) - B_1(\mathbf{x}).$$

If T_2^* were an estimator, it would be unbiased by construction. That is, $E[T_2^*(\mathbf{X})] = \max(\boldsymbol{\mu})$. However, since T_2^* contains the unknown $\boldsymbol{\mu}$, it is not computable. The bootstrap treats the sample as if it represents the bootstrap population, evaluating the bias as $E\left[T_1\left(\widetilde{\mathbf{X}}\right)\right] - \max(\mathbf{x})$, where $\widetilde{\mathbf{X}} \sim N\left(\mathbf{x}, \boldsymbol{\sigma}^2\right)$. Analytically, this is equivalent to replacing $B_1(\boldsymbol{\mu})$ with $B_1(\mathbf{x})$, so that T_2 is the bootstrap bias corrected estimator. Unfortunately, T_2 is not unbiased unless we have

$$E[B_1(\mathbf{X})] = B_1(\boldsymbol{\mu}).$$

To further analyze the bias, define the second-level bias function $B_2 \colon \mathbb{R}^j \to \mathbb{R}$ such that

$$B_2(\boldsymbol{\mu}) = E[T_2(\mathbf{X})] - \max(\boldsymbol{\mu}).$$

$B_2(\cdot)$ has the following property:

Fact 4.5. *$B_2(\boldsymbol{\mu}) < B_1(\boldsymbol{\mu})$, $\forall \boldsymbol{\mu} \in \mathbb{R}^j$.*

Proof. See appendix. □

Fact 4.5 justifies the usage of the bootstrap bias correction since the upward bias of T_1 will be reduced after the bootstrap correction. However, in general it cannot eliminate the bias. It is helpful to consider the case when $\mu_1 = ... = \mu_j$. As suggested in the proof of Proposition 4.4, $B_1(\boldsymbol{\mu})$ has already attained its maximum, while $E[B_1(\mathbf{X})]$ is the weighted average of $B_1(\cdot)$ evaluated at every realization of $\mathbf{X}$ with the weight given by the normal p.d.f. $\phi\left(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\sigma}^2\right)$. So we have $B_2(\boldsymbol{\mu}) = B_1(\boldsymbol{\mu}) - E[B_1(\mathbf{X})] > 0$. In that case, positive bias still exists after the bootstrap. Furthermore, it is possible that the bootstrap over-corrects the upward bias since $B_1(\boldsymbol{\mu})$ might be smaller than $E[B_1(\mathbf{X})]$ for some $\boldsymbol{\mu}$. For illustration, Figure 4.2 plots the two levels of bias functions when $j = 2$. We set $\sigma_1^2 = 1$, $\sigma_2^2 = 1$. Since only the difference between μ_1 and μ_2 matters, we normalize $\mu_1 = 0$ and plot B_1, B_2 against different values of μ_2. As we can see, i) when μ_2 goes to infinity or minus infinity, both $B_1$and B_2 approach zero; ii) the largest bias occurs when $\mu_2 = 0$; iii) the B_2 curve always lies below the B_1 curve; iv) though B_1 is always positive, there is a region that B_2 is slightly negative, which implies there is a possibility that the one-level bootstrap may over-correct the bias.

4.4.3 Multi-level bootstrap correction

Since one level of bootstrap estimator T_2 does not eliminate the bias, a natural extension is using its bias B_2 to further correct T_2. Define

$$T_3^*(\mathbf{x}) = T_2(\mathbf{x}) - B_2(\boldsymbol{\mu}),$$
$$T_3(\mathbf{x}) = T_2(\mathbf{x}) - B_2(\mathbf{x}).$$

Again, if T_3^* were an estimator, it would be unbiased by construction. However, our inability to evaluate $B_2(\cdot)$ at the right point, namely $\boldsymbol{\mu}$, forces us to compute $B_2(\mathbf{x})$ instead. In essence, we treat the sample $\mathbf{x}$ as the bootstrap population and evaluate $B_2(\mathbf{x}) = B_1(\mathbf{x}) - E\left[B_1\left(\widetilde{\mathbf{X}}\right)\right]$, where $\widetilde{\mathbf{X}} \sim N\left(\mathbf{x}, \boldsymbol{\sigma}^2\right)$. Since evaluating $B_1(\cdot)$ is equivalent to one level of bootstrap, evaluating $B_2(\cdot)$ can be viewed as doubling the bootstrap. Clearly, the estimator T_3 is not unbiased unless

we have

$$E\left[B_2\left(\mathbf{X}\right)\right] = B_2\left(\boldsymbol{\mu}\right).$$

The effect of bias reduction depends on the functional form of the bias function as well as the discrepancy between $\mathbf{x}$ and $\boldsymbol{\mu}$. The latter is unknown, and we cannot expect the realization $\mathbf{x}$ happens to be $\boldsymbol{\mu}$ in the finite sample. However, the bias function is under control in the sense that if $B_1\left(\cdot\right)$ were a linear function, T_2 would be unbiased regardless of the unknown $\boldsymbol{\mu}$. Similarly, if $B_2\left(\cdot\right)$ were a linear function, T_3 would be unbiased. We double the bootstrap because we hope $B_2\left(\cdot\right)$ ensembles more linearity. This raises two questions: Is $B_2\left(\cdot\right)$ flatter than $B_1\left(\cdot\right)$? If we proceed to higher level of the bootstrap, will we eventually obtain an unbiased estimator?

Define the higher-level bias function and bias corrected estimator as

$$\begin{aligned} B_i\left(\boldsymbol{\mu}\right) &= E\left[T_i\left(\mathbf{X}\right)\right] - \max\left(\boldsymbol{\mu}\right) \\ &= B_{i-1}\left(\boldsymbol{\mu}\right) - E\left[B_{i-1}\left(\mathbf{X}\right)\right], \\ T_{i+1}\left(\mathbf{x}\right) &= T_i\left(\mathbf{x}\right) - B_i\left(\mathbf{x}\right), \end{aligned}$$

for $i = 3, 4, 5, ...$

If we are willing to make an additional assumption, we have an answer to the above two questions.

Assumption 4.3. *$B_1\left(\boldsymbol{\mu}\right)$ can be well approximated by a polynomial.*

There is a need to justify this assumption. Note that $B_1\left(\boldsymbol{\mu}\right)$ is a continuous, but not differentiable function in that $\max\left(\cdot\right)$ is not differentiable. The Taylor theorem of polynomial approximation does not apply. However, in Eq. (4.2) and Eq. (4.4), $f_L\left(\mathbf{p}_i\right)$ is bounded by $[y_1, y_{n_Y}]$ and $[0, 1]$ respectively. Therefore, $\boldsymbol{\mu}$ is bounded. By the Stone-Weierstrass theorem, the bias function $B_1\left(\boldsymbol{\mu}\right)$ can be uniformly approximated by a polynomial.

Proposition 4.6. *Suppose $B_1\left(\boldsymbol{\mu}\right)$ is a polynomial of order d, where $d \geq 2$, then $B_2\left(\boldsymbol{\mu}\right)$ is a polynomial of order $d-2$. Each level of bootstrap will reduce the polynomial order by 2 successively. Bias can be eliminated after $\left[\frac{d}{2}\right]$ levels of bootstraps, where $[\cdot]$ refers to the operator of taking integers.*

Proof. See appendix. □

Let us illustrate this property with a numerical example. Consider two independent normal variates $X_i \sim N\left(\mu_i, \sigma_i^2\right)$, $i = 1, 2$. Assume $B_1(\boldsymbol{\mu}) = 2\mu_1^5\mu_2^6$, a polynomial of order 11.

$$\begin{aligned} E\left[B_1(\mathbf{X})\right] &= 2E\left(X_1^5\right) E\left(X_2^6\right) \\ &= 2\left(\mu_1^5 + 10\sigma_1^2\mu_1^3 + 15\sigma_1^4\mu_1\right) \cdot \left(\mu_2^6 + 15\sigma_2^2\mu_2^4 + 45\sigma_2^4\mu_2^2 + 15\sigma_2^6\right) \end{aligned}$$

When $B_1(\boldsymbol{\mu}) - E\left[B_1(\mathbf{X})\right]$, the leading term $2\mu_1^5\mu_2^6$ cancels, and there are no terms of order 10 like $\mu_1^5\mu_2^5$, $\mu_1^4\mu_2^6$. Therefore, $B_2(\boldsymbol{\mu})$ is reduced to a polynomial of order 9. If we forward the bootstrap to higher levels, then $B_3(\boldsymbol{\mu})$ will be a polynomial of order 7, and $B_4(\boldsymbol{\mu})$ of order 5, etc. Eventually $B_i(\boldsymbol{\mu})$ will be of order one or zero. $E\left[B_i(\mathbf{X})\right] = B_i(\boldsymbol{\mu})$ is satisfied, and $T_{i+1}(\mathbf{x})$ becomes an unbiased estimator. In other words, d rounds of the bootstraps can correct the bias for polynomial $B_1(\boldsymbol{\mu})$ of order up to $2d$.

4.4.4 Simultaneous bootstrap

The upper level bias function $B_i(\cdot)$ is constructed by the expectation of the lower level bias function $E\left[B_{i-1}(\cdot)\right]$, which has to be evaluated with simulation. The nested, iterative simulation suffers from the curse of dimensionality, and practically we are unable to proceed beyond double or triple bootstraps. To resolve the computational difficulty, we propose a simultaneous bootstrap algorithm which can conduct many level of bootstrap correction with affordable computational costs. Davidson and MacKinnon (2002, 2007) provide a similar procedure which they refer to as "fast double bootstrap".

The rationale for the simultaneous bootstrap comes from the identity

$$E_\xi\left\{E_{\eta|\xi}\left[g(\xi, \eta)\right]\right\} = E_{\xi,\eta}\left[g(\xi, \eta)\right],$$

for arbitrary random variables ξ, η and real valued function $g\colon \mathbb{R}^2 \to \mathbb{R}$, where the subscript in $E(\cdot)$ explicitly indicates random variables for which the expectation operator applies.

Suppose $E(\cdot)$ must be evaluated with simulation. The left hand side of that identity prescribes a nested procedure. In the first step we draw a ξ. Conditional on this value of ξ, we

draw thousands of η, and then average $g(\xi, \eta)$. In the second step, we repeat the first step with thousands of ξ, and then average the averaged $g(\xi, \eta)$. However, the right hand side prescribes a simultaneous procedure such that we draw (ξ, η) from their joint distribution, and take the average of $g(\xi, \eta)$.

Given the same computational costs measured as the number of visits to $g(\xi, \eta)$, the latter procedure provides a more accurate approximation. This is because in the simultaneous simulation procedure draws of the pair (ξ, η) are independent, while in the nested simulation the same draw of ξ needs to be used for multiple times, which induces positive correlation and larger variance. To formalize this idea, we present the following proposition.

Proposition 4.7 (Efficiency of simultaneous simulation)**.** *Let the simulator for* $E_{\xi,\eta}[g(\xi, \eta)]$ *be*

$$S_1 = \frac{1}{N^2} \sum_{i=1}^{N^2} g(\xi_i, \eta_i),$$

where $\{\xi_i, \eta_i\}_{i=1}^{N^2}$ *are i.i.d. draws from the joint distribution of* (ξ, η)*.*

Let the simulator for $E_\xi \left\{ E_{\eta|\xi}[g(\xi, \eta)] \right\}$ *be*

$$S_2 = \frac{1}{N} \sum_{j=1}^{N} \left[\frac{1}{N} \sum_{k=1}^{N} g(\xi_j, \eta_{j,k}) \right],$$

where $\{\xi_j\}_{j=1}^{N}$ *are i.i.d. draws from the marginal distribution of* ξ*, while* $\{\eta_{j,k}\}_{k=1}^{N}$ *are i.i.d. draws from the conditional distribution of* $\eta | (\xi = \xi_j)$*,* $j = 1, ..., N$*.*

Then we have

$$E(S_1) = E(S_2),$$

$$Var(S_1) \leq Var(S_2),$$

with equality of variance iff $E_{\eta|\xi}[g(\xi, \eta)] = E_{\xi,\eta}[g(\xi, \eta)]$ *for all realizations of* ξ*.*

Proof. See appendix. □

To illustrate the efficiency of the simultaneous simulation relative to the nested simulation, consider a simple numerical example.

Let $(\xi, \eta) \sim N\left(0, 0, 1^2, 1^2, 0.5\right)$, $g(\xi, \eta) = \xi + \eta$, $N = 10$.

Then $Var(S_1) = Var\left[\frac{1}{100}\sum_{i=1}^{100}(\xi_i + \eta_i)\right] = \frac{3}{100}$,
but $Var(S_2) = Var\left[\frac{1}{100}\sum_{j=1}^{10}\sum_{k=1}^{10}(\xi_j + \eta_{j,k})\right] = \frac{21}{100}$.

We see that the nested simulation has a variance seven times larger than the simultaneous procedure, given 100 visits to $g(\xi, \eta)$ in both procedures. Even if we change the correlation of (ξ, η) from 0.5 to 0, nested simulation still has a larger variance. In that case, we have $Var(S_1) = \frac{2}{100}$, and $Var(S_2) = \frac{11}{100}$. The inflation of variance is due to the fact that the same draw of ξ_j has to be used 10 times in nested simulation.

Generally speaking, the simultaneous simulation will substantially improve the quality of the simulator. The only case of no improvement is that the conditional expectation is identical to the unconditional expectation for all realizations of the variable being conditioned on. To give a example, consider $(\xi, \eta) \sim N\left(0, 0, 1^2, 1^2, 0.5\right)$ with $g(\xi, \eta) = \xi\eta$. In that case, $Var(S_1) = Var(S_2) = \frac{5}{1000}$. However, once (ξ, η) have non-zero means, there will be improvement.

The results can be extended to multivariate and vector-valued random variables. We have the identity

$$E_{\xi_1} E_{\xi_2|\xi_1} \ldots E_{\xi_n|\xi_{n-1}\ldots\xi_1} g(\xi_1, ..., \xi_n) = E_{\xi_1,\ldots,\xi_n}\left[g(\xi_1, ..., \xi_n)\right],$$

for arbitrary vector-valued random variables $\xi_1, ..., \xi_n$ and real valued function g.

Again, the left hand side prescribes a multi-level nested simulation procedure, while the right hand side suggests a simultaneous simulation algorithm. The inefficiency of the nested procedure comes from the multiple usage of the same draw of ξ_{n-1}, and of ξ_{n-2}, ..., and worst of all, of ξ_1.

Multi-level bootstrap bias correction is a direct application of the above results.

Though $B_1(\cdot)$ might be evaluated by analytic formula or deterministic quadrature, $B_2(\cdot)$, $B_3(\cdot)$, etc. need to be evaluated by simulation. For example, consider evaluating $B_3(\mathbf{x})$:

$$\begin{aligned}
B_3(\mathbf{x}) &= B_2(\mathbf{x}) - E_{\mathbf{X}} B_2(\mathbf{X}) \\
&= \left[B_1(\mathbf{x}) - E_{\mathbf{X}} B_1(\mathbf{X})\right] - E_{\mathbf{X}}\left[B_1(\mathbf{X}) - E_{\widetilde{\mathbf{X}}|\mathbf{X}} B_1\left(\widetilde{\mathbf{X}}\right)\right] \\
&= E_{\mathbf{X}} E_{\widetilde{\mathbf{X}}|\mathbf{X}}\left\{\left[B_1(\mathbf{x}) - B_1(\mathbf{X})\right] - \left[B_1(\mathbf{X}) - B_1\left(\widetilde{\mathbf{X}}\right)\right]\right\} \\
&= E_{\mathbf{X},\widetilde{\mathbf{X}}}\left[g\left(\mathbf{X},\widetilde{\mathbf{X}}\right)\right]
\end{aligned}$$

where $\mathbf{X} \sim N\left(\mathbf{x}, \boldsymbol{\sigma}^2\right)$, $\widetilde{\mathbf{X}}\left|(\mathbf{X}=\mathbf{y}) \sim N\left(\mathbf{y}, \sigma^2\right)\right.$. $g\left(\mathbf{X}, \widetilde{\mathbf{X}}\right) = \left[B_1(\mathbf{x}) - B_1(\mathbf{X})\right] - \left[B_1(\mathbf{X}) - B_1\left(\widetilde{\mathbf{X}}\right)\right]$.

The simultaneous procedure for $B_3(\mathbf{x})$ takes the following steps:

First, sample a pair $(\mathbf{y}, \mathbf{z})$ from the joint distribution of $\left(\mathbf{X}, \widetilde{\mathbf{X}}\right)$. The easiest way is the method of composition, that is, to sample $\mathbf{y}$ from $N\left(\mathbf{x}, \sigma^2\right)$, and then sample $\mathbf{z}$ from $N\left(\mathbf{y}, \boldsymbol{\sigma}^2\right)$.

Second, evaluate $g(\mathbf{y}, \mathbf{z})$, which is a difference of differenced $B_1(\cdot)$.

Third, repeat the first and second step, and average the results.

A higher order bias function $B_i(\cdot)$, $i > 3$ can be simultaneously simulated in the same way. The first step is a hierarchical sampling of normal variates. The second step is a multiple difference of $B_1(\cdot)$ evaluated at the obtained sample.

From the perspective of computation, instead of being evaluated directly, $B_1(\cdot)$ may be treated as another level (that is, the bottom level) of the simultaneous simulation. It is less precise, but much faster. The saved computation time can be used for a larger scale simulation, which improves the precision of all levels of bootstraps. Given the same computation costs measured in CPU time, whether the gains outweighs the loss is a practical issue.

4.5 Monte Carlo evidence

In this section, we replicate the Monte Carlo experiment in Manski and Pepper (2009), with multi-level bootstrap added to further reduce the bias. The experiment simulates the MIV lower bound of the treatment response $E(Y_t | Z = z_j)$ as in Eq. (4.1). The joint distribution of (Y, D, Z) is specified in the identical way as in Manski and Pepper (2009). The MIV Z has a categorical distribution with M equal-probability mass points $\left\{\frac{1}{M}, \frac{2}{M}, ..., 1\right\}$. The treatment variable $D = I(Z + \varepsilon > 0)$, where $\varepsilon \sim N(0, 1)$. The response variable Y follows $N\left(0, \sigma^2\right)$ censored to $(-1.96, 1.96)$. With a random sample of n observations, we evaluate the Monte Carlo distribution of the analogue MIV bound for $E(Y_1 | Z = 1)$ with 1000 repetitions.

Our bootstrap correction algorithm assumes normality as well as fixed variances. The finite-sample variances are computed from the analogue version of the formula in Proposition 4.3. Note that there is no need to discretize Y when we apply that formula since analogue conditional variance can be used. This is advantageous to the asymptotic variances given by Proposition

4.2, where we have to discretize every variable and calculate the gradients. Nevertheless, the computed variances are close regardless of the approach.

Once we obtain the variances, we apply the simultaneous multi-level bootstrap procedure to correct the bias. Simulation results of the first four levels of bootstraps on the basis of 100000 draws are presented in Table 4.1. Each column is an experiment with selected values of M, σ^2, n. The fourth row displays the biases of the raw analogue estimator (T_1), which are comparable to Table 1 in Manski and Pepper (2009). The fifth row shows the biases of first-level bootstrap corrected estimator (T_2), comparable to their Table 2 in Manski and Pepper (2009). The following rows show the biases of second, third, fourth levels of bootstrap corrected estimators (T_3, T_4, T_5). The last row presents the biases of the conservative estimator (T_c), which is biased downwards.

Our results of the biases of T_1 and T_2 are very close to those reported by Manski and Pepper (2009). The slight difference might due to the fact that they used nonparametric bootstrap (resampling from the empirical distribution) while we use parametric bootstrap (resampling from the normal distribution with estimated variance). The most important new result is that T_3, T_4, T_5 have smaller biases. For example, in the setting $M = 8$, $\sigma^2 = 25$, $n = 100$, the analogue estimator T_1 has a huge bias of 0.55. The first level bootstrap reduces the bias to 0.22, but the bias is still relatively large. As predicted by Proposition 4.6, higher level of bootstrapped estimator T_3, T_4, T_5 further improve the estimator with biases of 0.15, 0.11, 0.09 respectively. In fact, in most M, σ^2, n settings the simulated biases are monotone decreasing as the bootstrap is forwarded to a higher level.

Also note that when the bias has already achieved a tiny level (compared to the numerical standard errors of simulation), a further bootstrap may not improve the estimator any more, but there is also no sign of deterioration. This observation is in line with Proposition 4.6, which indicates that d rounds of bootstraps can correct the bias for polynomial $B_1(\boldsymbol{\mu})$ of order up to $2d$. After that, the bias function becomes a constant with no improvement afterwards. This happens mostly in settings where $n = 1000$. In those cases, since the raw analogue estimator is consistent, the finite sample bias of T_1 is already small. We cannot expect that a multi-level bootstrap will eliminate the bias because high dimensional simulation itself introduces non-

negligible error. As a practical suggestion, we recommend more levels of bootstrap correction when the sample size is small, but one or two levels of bootstrap may suffice for a large dataset. Of course, increasing simulation draws will make higher level bootstrap bias correction more reliable, if we can afford the computation costs.

The simulation results also suggest the usefulness of the conservative estimator T_c. If we prefer some wider bounds than the true bounds and are not willing to resort to any bootstrap correction, we may use the conservative estimator. For $M = 4$, the magnitude of downward bias induced by T_c is relatively larger than the magnitude of upward bias caused by T_1, though still on the same scale. For $M = 8$, the absolute size of bias are similar between T_c and T_1. Furthermore, as n becomes larger, T_c decreases as well, which suggests that in large sample T_c offers a cheap but effective solution to the problematic analogue MIV bounds.

4.6 An application to disability misreporting identification

In this section, we reconsider the empirical study of Kreider and Pepper (2007) on the employment gap between disabled and non-disabled persons. The employment gap is defined as

$$\begin{aligned} &P(L=1\,|W=1) - P(L=1\,|W=0) \\ &= \sum P(Z=z_j) \cdot [P(L=1\,|W=1, Z=z_j) - P(L=1\,|W=0, Z=z_j)], \end{aligned}$$

where the MIV bounds of $P(L=1\,|W=1, Z=z_j)$ is given by Eq. (4.4), and the bounds of $P(L=1\,|W=0, Z=z_j)$ can be formulated similarly.

Kreider and Pepper (2007) analyze two datasets: 1992-93 Health and Retirement Study (HRS) and 1996 Survey of Income and Program Participation (SIPP) with sample sizes 12,503 and 29,807 respectively. Respondents' employment status (L), reported disability status (X) and grouped age (Z) can be directly observed in the data. As for the verification status (Y), it depends on how researchers use prior information to classify the verified group. They consider five different ways to define the verified subpopulation: a) disability beneficiaries; b) those verified in Wave 2; c) gainfully employed workers; d) those claiming no disability in the current

wave; e) all of the above. Readers are referred to Kreider and Pepper (2007, p.435) for the detailed definition of subgroups.

From the data, the analogue joint probability of (L, X, Y, Z) is obtained, and then the analogue bounds of employment gap are computed. Then we use simultaneous multi-level bootstraps to correct the biases. The estimated bounds are presented in Table 4.2. Our results on the raw analogue bounds and first-level bootstrap corrected bounds (T_1 and T_2) are almost identical to those reported by Kreider and Pepper (2007) in their Table 4, despite that they used the standard non-parametric bootstrap and we use the normal distribution with estimated variances to correct the biases. This is because the current sample size is large, and the estimated probability vector is well approximated by the multivariate normal variates. As a result, our parametric bootstrap works well.

In the finite sample, the raw analogue bounds are narrower than the true bounds on average. After the bootstrap correction, the bounds become wider. It seems that the first-level bootstrap does not fully remove the bias since higher order bootstraps further widen the estimated bounds. This is most apparent for the HRS data. For example, in the beneficiaries verification scenario the analogue bounds are $[-0.959, 0.809]$, the first-level bootstrap magnifies the bounds to $[-0.971, 0.830]$, and further bootstraps expand them to $[-0.975, 0.836]$ and $[-0.978, 0.839]$, and so on. Of course the speed of expansion decreases with the level of bootstraps. As an empirical guide, when the expansion mitigates, it is better to stop increasing the bootstrap levels. For the SIPP data, the sample size is twice as large as that of the HRS data. Therefore, the speed of bounds expansion is modest. It seems that one or two levels of bootstraps suffices to remove most of the biases.

It is worth mentioning that the conservative estimator T_c provides the widest bounds. This is not surprising since the conservative lower (upper) bound is biased downwards (upwards). However, it is not too wide to be informative. Whenever the raw analogue bounds and bootstrap corrected bounds are indecisive on the sign of the employment gap, so are conservative bounds. Only in the last case, the analogue estimator indicates that the employment gap in the SIPP data is negative and bounded by $[-0.413, -0.224]$. Three levels of bootstraps widen the bounds to $[-0.447, -0.199]$, and the conservative estimator also suggests the gap is negative

and bounded by $[-0.482, -0.131]$.

4.7 Conclusion

In practice, the MIV assumption is useful in partial identification of treatment effects in that an MIV is easier to provide and justify than a standard instrumental variable. Implementation of an MIV consists of two steps. First, we obtain the worse-case bounds conditional on each value of an MIV. Second, the maximum (minimum) of the worst-case bounds corresponding to the smaller (larger) MIV values than the current MIV value yields sharp lower (upper) MIV bounds on the treatment effect. Unfortunately, as noted by Manski and Pepper (2000), the maximum and minimum operators make the analogue bounds narrower than the true bounds in finite samples. This problem is more acute when i) the worse-case bounds under each value of an MIV are close to each other; ii) the variances of analogue bounds are large; and iii) the number of discrete MIV values is large. In such cases, it is sensible to adjust the analogue estimator so as to avoid over-optimistic inference.

This paper provides two new types of analogue estimator adjustment. The first derives a conservative estimator that is obtained by subtracting the largest possible bias from the analogue estimator. This approach is most useful when the sample size is large so that the variances of the analogue bounds are small. In that case, the largest possible bias is small, and therefore the conservative bounds are likely to remain informative. Another virtue of the conservative estimator is its inexpensiveness of computation.

The second solution introduces a computationally feasible multi-level bootstrap correction. It is shown that one level of the bootstrap cannot eliminate the bias in general, and there is also a possibility of over-correction. This inadequacy of the single bootstrap leaves room for higher level bootstraps, which do not necessarily suffer from the curse of dimensionality in that a simultaneous simulation strategy can make multi-level bootstraps computationally feasible. From a practical side, a practitioner simply inputs the analogue worse-case bounds and standard errors for each value of the MIV, and the simulation outputs include the corrected estimators after each level of bootstrap. From a theoretical perspective, under an assumption that the bias function can be well approximated by a polynomial, each level of bootstrap is

shown to successively reduce the polynomial order of the bias function. Monte Carlo evidence provides strong support for the effectiveness of the multi-level bootstrap correction.

n	100	100	100	100	100	100
M	4	4	4	8	8	8
σ^2	1	4	25	1	4	25
T1	0.10	0.15	0.20	0.31	0.42	0.53
T2	0.01	0.03	0.06	0.09	0.14	0.21
T3	0.00	0.01	0.03	0.04	0.07	0.13
T4	-0.01	0.00	0.02	0.02	0.03	0.09
T5	-0.01	-0.01	0.01	0.00	0.01	0.07
Tc	-0.15	-0.16	-0.17	-0.22	-0.23	-0.23
n	500	500	500	500	500	500
M	4	4	4	8	8	8
σ^2	1	4	25	1	4	25
T1	0.02	0.02	0.04	0.08	0.12	0.15
T2	0.00	-0.01	-0.01	0.01	0.03	0.04
T3	-0.01	-0.02	-0.02	0.00	0.01	0.01
T4	-0.01	-0.02	-0.02	0.00	0.00	0.00
T5	-0.01	-0.02	-0.02	-0.01	0.00	0.00
Tc	-0.09	-0.11	-0.12	-0.14	-0.15	-0.16
n	1000	1000	1000	1000	1000	1000
M	4	4	4	8	8	8
σ^2	1	4	25	1	4	25
T1	0.00	0.01	0.02	0.04	0.07	0.09
T2	-0.01	-0.01	0.00	0.00	0.01	0.02
T3	0.00	-0.01	0.00	0.00	0.01	0.01
T4	0.00	-0.01	0.00	-0.01	0.00	0.01
T5	0.00	-0.01	0.00	-0.01	0.00	0.01
Tc	-0.07	-0.09	-0.09	-0.11	-0.12	-0.13

T1 is the average bias of the naive estimator (maximum of the sample). T2 is the average bias of first-level bootstrap corrected estimator. T3, T4, T5 are biases of second-, third-, fourth- level bootstrap corrected estimators. Tc is the bias of the (downward biased) conservative estimator. Two decimals are retained since the average numerical standard error is 0.007 (maximum 0.022, minimum 0.002)

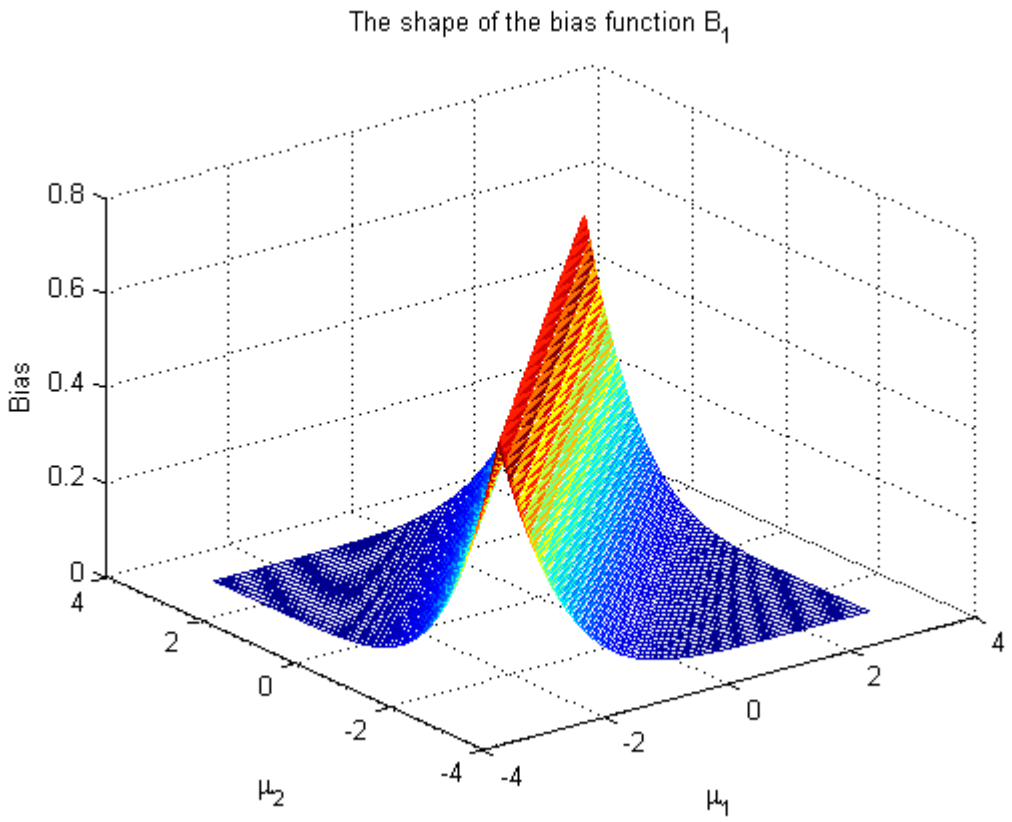
Table 4.1 Bias of analogue estimate of the MIV lower bound with the bootstrap correction

HRS	Beneficiaries	Wave 2	Workers	No disability	All of above
T1	[-0.959, 0.809]	[-0.741, 0.645]	[-0.811, 0.350]	[-0.760, 0.350]	[-0.402, -0.341]
T2	[-0.971, 0.830]	[-0.760, 0.672]	[-0.824, 0.358]	[-0.767, 0.358]	[-0.430, -0.307]
T3	[-0.975, 0.836]	[-0.763, 0.681]	[-0.826, 0.359]	[-0.766, 0.359]	[-0.434, -0.302]
T4	[-0.978, 0.839]	[-0.764, 0.688]	[-0.826, 0.359]	[-0.766, 0.359]	[-0.434, -0.300]
Tc	[-0.980, 0.857]	[-0.794, 0.704]	[-0.847, 0.383]	[-0.788, 0.383]	[-0.492, -0.217]

SIPP	Beneficiaries	Wave 2	Workers	No disability	All of above
T1	[-0.967, 0.908]	[-0.793, 0.869]	[-0.784, 0.318]	[-0.781, 0.318]	[-0.413, -0.224]
T2	[-0.974, 0.915]	[-0.804, 0.880]	[-0.794, 0.322]	[-0.785, 0.322]	[-0.437, -0.202]
T3	[-0.977, 0.916]	[-0.808, 0.882]	[-0.795, 0.322]	[-0.786, 0.322]	[-0.444, -0.199]
T4	[-0.978, 0.917]	[-0.811, 0.883]	[-0.795, 0.322]	[-0.786, 0.322]	[-0.447, -0.199]
Tc	[-0.982, 0.925]	[-0.820, 0.900]	[-0.816, 0.346]	[-0.797, 0.346]	[-0.482, -0.131]

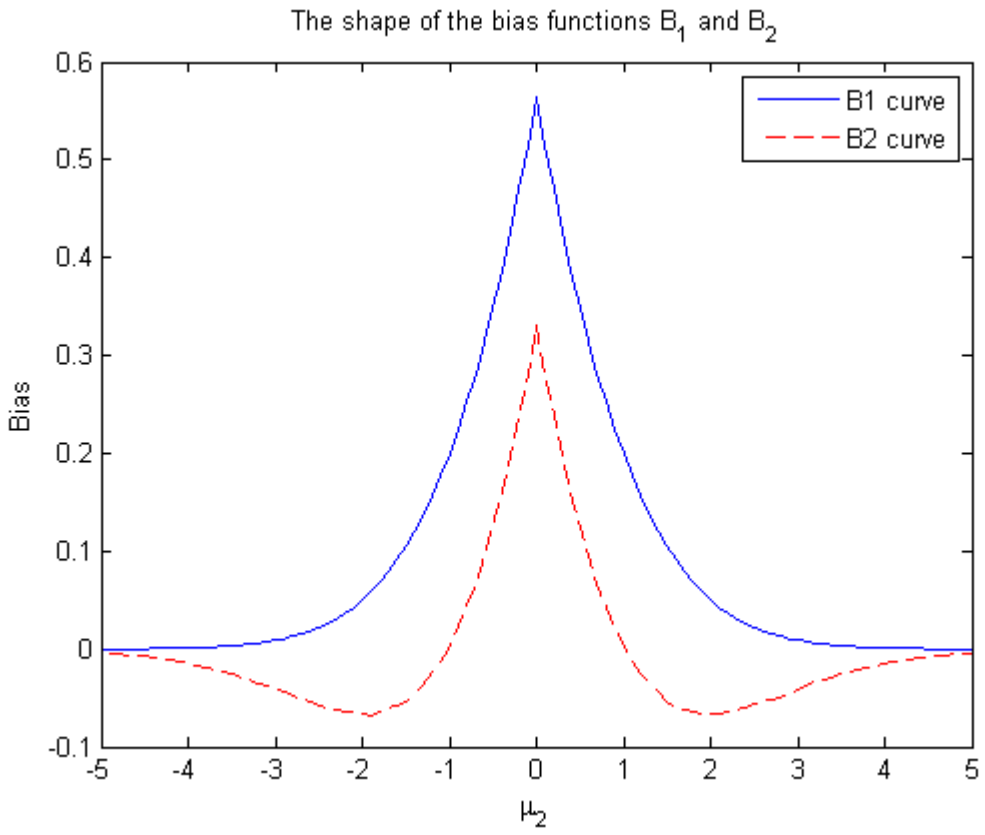
Beneficiaries, Wave 2, Workers, No disability are defined identically as in Kreider and Pepper (2007). T1 is the raw analogue estimator,that is,maximum of the sample, comparable to Table 4 in Kreider and Pepper (2007). T2 is first-level bootstrap corrected estimator, comparable to Table 4 in Kreider and Pepper (2007). T3 is second-level bootstrap corrected estimator. T4 is the third-level bootstrap corrected estimator. The upper panel shows the results for the HRS dataset, and the lower panel for SIPP dataset.

Table 4.2 MIV bounds of employment gap with the bootstrap correction



The first-level bias (B_1) is plotted for the case of two normal variates. The two arguments of B_1 function is the mean of the two normal variates. We set $\sigma_1^2 = 1$, $\sigma_2^2 = 1$.

Figure 4.1 The shape of the bias function after the first level bootstrap



The first level (B_1) and second level (B_2) of the bias functions are plotted for the case of two normal variates. We set $\sigma_1^2 = 1$, $\sigma_2^2 = 1$. Since only the difference in mean matters, μ_1 is normalized to zero. As μ_2 moves, the magnitude of the first-level bias and the second-level bias change accordingly. However, the B_1 curve always lies above the B_2 curve. Though B_1 is always positive, there is a region where B_2 falls below zero.

Figure 4.2 The shape of the bias functions after two levels of bootstrap

CHAPTER 5. GENERAL CONCLUSION AND DISCUSSION

Missing data is a common problem, making it difficult to draw an accurate and complete picture of the economy. However, there are efficient methods handling imperfectly observed data if its generating process and missing mechanism is known to the researcher. In this dissertation, temporal and cross-sectional aggregations are framed in two parametric models, namely a VAR model in Chapter 2 and a two-equation imputation model in Chapter 3. There are minor differences between them. First, the former is more symmetric with respective to regressors than the latter. Second, the former has autocorrelations in variables while the latter does not. Apart from these differences on the surface, the two models are identical in the sense that we use the same idea to handle the data aggregation. In summary, we solve the problem by two steps. First, explore the joint distribution of all disaggregated variables. Second, marginalize unobserved variables out of the joint distribution. In the first step, disaggregated variables have their own joint distribution depending on the model structure. The two-equation imputation model does not have autocorrelations in variables, so the covariance matrix is block-diagonal. The VAR model has both autocorrelations and crosscorrelations among variables, but the covariance matrix is still straightforward to derive for a stationary VAR process. In the second step, the marginalization is tractable for multivariate normal random variables, for the observed aggregated variables are simply the linear combinations of these disaggregated variables. As a result, we obtain the explicit likelihood function corresponding to the observed data. No matter whether the inference is based on the classical maximum likelihood or on the Bayesian data augmentation simulation, we mainly use the information conveyed by the joint distribution of the observed data.

One concern is departure from normality. In that circumstance, to marginalize the latent disaggregated variables out of the likelihood function may involve an integral that does not lead

to a closed-form solution. However, if the underlying distribution is finite Gaussian mixtures, Markov switching Gaussian distributions, scale mixtures such as student-t, the problem is still tractable. In these cases, normality is not completely lost in the sense that conditional on latent regimes/states, the joint distribution of disaggregated variables is still multivariate normal. The Bayesian approach will have an advantage addressing those complications in that the Gibbs sampler handles posterior disaggregated variables by conditioning on all the other model parameters and latent components in the model.

Another concern is the model uncertainty. The validity of our methods relies on the correct specification of the probabilistic model. A right model is crucial for an insightful analysis of temporal or cross-sectional aggregation issues. In reality, a researcher often does not know which model best describes the data generating process of the disaggregated variables. Uncertainties on the model can be handled in the parametric framework through the Bayesian model selection or averaging. Before an empirical study, a researcher is likely to conceive candidate models that plausibly generate the data. After observing the data, the likelihood function under each candidate model can serve as the criterion for model selection or the weight for the model averaging. Note that our method of handling data aggregation can be applied to many types of regression-type models. In principle, the method can be applied to each candidate model to find the likelihood conditional on the model. Then an eclectic inference or prediction can be made through the model averaging.

Chapter 4 discusses another type of missing data due to the counterfactual outcomes. We propose a multi-level bootstrap solution to the finite sample bias caused by the MIV identification of the counterfactuals. Our contribution is that the mechanism of bias correction is not based on asymptotic refinement of the bootstrapped estimator, but on a direct comparison of the bias functions before and after the bootstrap correction. The usefulness of our approach is supported by Monte Carlo studies. However, more justifications, or perhaps relaxation, on the polynomial approximation to the bias function are left for future research.

APPENDIX A. ADDITIONAL MATERIAL FOR CHAPTER 2

A.1 Proof of Proposition 2.1

Let $\mathbf{Z} \equiv \begin{pmatrix} \mathbf{Z}_p^* \\ \mathbf{Z}_{p+1}^* \\ \cdots \\ \mathbf{Z}_T^* \end{pmatrix}$, $\widetilde{\mathbf{Z}} \equiv \begin{pmatrix} \mathbf{Z}_p^* \\ \mathbf{Z}_{p+1}^* - \mathbf{F}\mathbf{Z}_p^* \\ \cdots \\ \mathbf{Z}_T^* - \mathbf{F}\mathbf{Z}_{T-1}^* \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_p^* \\ \mathbf{e}_{p+1} \\ \cdots \\ \mathbf{e}_T \end{pmatrix}$.

By the normality and independence assumption of $\mathbf{Z}_p^*, \mathbf{e}_{p+1}, ..., \mathbf{e}_T$, we have $\widetilde{\mathbf{Z}} \sim N\left(\mathbf{0}, \widetilde{\mathbf{\Delta}}\right)$, where

$$\widetilde{\mathbf{\Delta}} = \begin{pmatrix} \mathbf{B} & & & \\ & \mathbf{\Delta} & & \\ & & \ddots & \\ & & & \mathbf{\Delta} \end{pmatrix}.$$

Construct a $kp\left(T-p+1\right)$-by-$kp\left(T-p+1\right)$ matrix $\mathbf{L} = \begin{pmatrix} \mathbf{I} & & & \\ -\mathbf{F} & \mathbf{I} & & \\ & \ddots & & \\ & & -\mathbf{F} & \mathbf{I} \end{pmatrix}$, and its inverse is given by $\mathbf{L}^{-1} = \begin{pmatrix} \mathbf{I} & & & \\ \mathbf{F} & \mathbf{I} & & \\ \cdots & & \ddots & \\ \mathbf{F}^{T-p} & \mathbf{F}^{T-p-1} & \mathbf{F} & \mathbf{I} \end{pmatrix}$. Clearly $\widetilde{\mathbf{Z}} = \mathbf{L}\mathbf{Z}$. It follows that $\mathbf{Z} = \mathbf{L}^{-1}\widetilde{\mathbf{Z}}$ and $\mathbf{Z} \sim N\left[\mathbf{0}, \left(\mathbf{L}^{-1}\right)\widetilde{\mathbf{\Delta}}\left(\mathbf{L}^{-1}\right)'\right]$. After a little algebra and repeated use of the

identity $\mathbf{B} = \mathbf{F}\mathbf{B}\mathbf{F}' + \mathbf{\Delta}$, we have

$$\left(\mathbf{L}^{-1}\right)\widetilde{\mathbf{\Delta}}\left(\mathbf{L}^{-1}\right)' = \begin{pmatrix} \mathbf{B} & (\mathbf{F}\mathbf{B})' & \cdots & \left(\mathbf{F}^{T-p}\mathbf{B}\right)' \\ \mathbf{F}\mathbf{B} & \mathbf{B} & \cdots & \left(\mathbf{F}^{T-p-1}\mathbf{B}\right)' \\ \vdots & & \ddots & \\ \mathbf{F}^{T-p}\mathbf{B} & \mathbf{F}^{T-p-1}\mathbf{B} & \cdots & \mathbf{B} \end{pmatrix}.$$

The structure of $\left(\mathbf{L}^{-1}\right)\widetilde{\mathbf{\Delta}}\left(\mathbf{L}^{-1}\right)'$ implies that the joint distribution of $\mathbf{Z}_t^*, ..., \mathbf{Z}_{t+s}^*$ is identical to that of $\mathbf{Z}_{t+j}^*, ..., \mathbf{Z}_{t+s+j}^*$ for all t, s, j, hence the strict stationarity of $\{\mathbf{Z}_t^*\}_{t=p}^T$.

Furthermore, $E\left(\mathbf{Z}_t^*\mathbf{Z}_{t-1}^{*\prime}\right) = \mathbf{F}\mathbf{B}$, and $\mathbf{\Gamma}_p = E\left[\left(\mathbf{Y}_t^* - \boldsymbol{\mu}_1\right)\left(\mathbf{Y}_{t-p}^* - \boldsymbol{\mu}_1\right)'\right]$ is the submatrix of $\mathbf{F}\mathbf{B}$ selected by its first k rows and last k columns. When eigenvalues of $\mathbf{F}$ all lie inside the unit circle, $\mathbf{B}$ is a well-defined covariance matrix and takes the form

$$\mathbf{B} = \begin{pmatrix} \mathbf{\Gamma}_0 & \mathbf{\Gamma}_1 & ... & \mathbf{\Gamma}_{p-1} \\ \mathbf{\Gamma}_1' & \mathbf{\Gamma}_0 & ... & \mathbf{\Gamma}_{p-2} \\ ... & & & \\ \mathbf{\Gamma}_{p-1}' & \mathbf{\Gamma}_{p-2}' & ... & \mathbf{\Gamma}_0 \end{pmatrix}.$$

It follows that $\mathbf{\Gamma}_p = \sum_{i=1}^p \mathbf{\Phi}_i\mathbf{\Gamma}_{p-i}$.

Similarly, $\mathbf{\Gamma}_{p+1}$ is the submatrix of $\mathbf{F}^2\mathbf{B}$ by extracting its first k rows and last k columns. The first k rows of $\mathbf{F}$ is $(\mathbf{\Phi}_1, ..., \mathbf{\Phi}_p)$ and the last k columns of $\mathbf{F}\mathbf{B}$ is $\left(\mathbf{\Gamma}_p', ..., \mathbf{\Gamma}_1'\right)'$. Their inner product is $\mathbf{\Gamma}_{p+1} = \sum_{i=1}^p \mathbf{\Phi}_i\mathbf{\Gamma}_{p+1-i}$. By induction, $\mathbf{\Gamma}_j$ is the submatrix of $\mathbf{F}^{j-p+1}\mathbf{B}$ by extracting its first k rows and last k columns, which is the inner product of the first k rows of $\mathbf{F}$ and the last k columns of $\mathbf{F}^{j-p}\mathbf{B}$. So we have $\mathbf{\Gamma}_j = \sum_{i=1}^p \mathbf{\Phi}_i\mathbf{\Gamma}_{j-i}$.

A.2 Proof of Proposition 2.2

For notational convenience, define $\mathbf{Y}_s^t = \{\mathbf{Y}_s^*, ..., \mathbf{Y}_t^*\}$. Let

$$\mathbf{Y}_t^* = \mathbf{c} + \sum_{i=1}^p \mathbf{\Phi}_i\mathbf{Y}_{t-i}^* + \varepsilon_t.$$

This data generating process suggests

$$p\left(\mathbf{Y}_t^* \left| \mathbf{Y}_{t-p}^{t-1}\right.\right) = p\left(\mathbf{Y}_t^* \left| \mathbf{Y}_{t-p-1}^{t-1}\right.\right) = ... = p\left(\mathbf{Y}_t^* \left| \mathbf{Y}_1^{t-1}\right.\right),$$

since their distributions are all equal to the distribution of ε_t with the mean shifted by $\mathbf{c} + \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{Y}_{t-i}^*$. Then we have

$$\begin{aligned} p\left(\mathbf{Y}_s^t \left| \mathbf{Y}_1^{s-1}, \mathbf{Y}_{t+1}^T \right.\right) &\propto p\left(\mathbf{Y}_1^T\right) \\ &= p\left(\mathbf{Y}_1^{s-1}\right) \cdot \prod_{j=s}^{t+p} p\left(\mathbf{Y}_j^* \left| \mathbf{Y}_{j-p}^{j-1} \right.\right) \cdot p\left(\mathbf{Y}_{t+p+1}^T \left| \mathbf{Y}_{t+1}^{t+p} \right.\right) \\ &\propto \prod_{j=s}^{t+p} p\left(\mathbf{Y}_j^* \left| \mathbf{Y}_{j-p}^{j-1} \right.\right). \end{aligned}$$

Similarly,

$$\begin{aligned} p\left(\mathbf{Y}_s^t \left| \mathbf{Y}_{s-p}^{s-1}, \mathbf{Y}_{t+1}^{t+p} \right.\right) &\propto p\left(\mathbf{Y}_{s-p}^{t+p}\right) \\ &= p\left(\mathbf{Y}_{s-p}^{s-1}\right) \cdot \prod_{j=s}^{t+p} p\left(\mathbf{Y}_j^* \left| \mathbf{Y}_{j-p}^{j-1} \right.\right) \\ &\propto \prod_{j=s}^{t+p} p\left(\mathbf{Y}_j^* \left| \mathbf{Y}_{j-p}^{j-1} \right.\right). \end{aligned}$$

Both $p\left(\mathbf{Y}_s^t \left| \mathbf{Y}_1^{s-1}, \mathbf{Y}_{t+1}^T \right.\right)$ and $p\left(\mathbf{Y}_s^t \left| \mathbf{Y}_{s-p}^{s-1}, \mathbf{Y}_{t+1}^{t+p} \right.\right)$ are proper. If they are proportional to the same expression, they must be equal.

A.3 The State space form of the varied frequency VAR

Let k dimensional latent $\{\mathbf{Y}_t^*\}_{t=1}^T$ follow a stationary VAR(p) process:

$$\mathbf{Y}_t^* = \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{Y}_{t-i}^* + \varepsilon_t,$$

where $\varepsilon_t \sim N(\mathbf{0}, \mathbf{\Omega})$. For simplicity, consider the case of balanced temporal aggregation. Partition $\mathbf{Y}_t^*$ into $\mathbf{Y}_{1,t}^*$ and $\mathbf{Y}_{2,t}^*$, where $\mathbf{Y}_{1,t}^*$ is fully observed but $\mathbf{Y}_{2,t}^*$ is only observed every q period. In other words, $\mathbf{Y}_{1,t} = \mathbf{Y}_{1,t}^*$, for $t = 1, ..., T$, but $\mathbf{Y}_{2,t} = \sum_{j=0}^{q-1} \mathbf{Y}_{2,t-j}^*$, for $t = q, 2q, 3q, ..., T$. To write the system into the state space form, we keep track of $r \equiv \max(p, q)$ recent periods of $\mathbf{Y}_t^*$ as the state vector. Let $\boldsymbol{\xi}_t = \left(\mathbf{Y}_t^{*\prime}, ..., \mathbf{Y}_{t-r+1}^{*\prime}\right)'$, $\boldsymbol{F} = \begin{pmatrix} \mathbf{\Phi} \\ \mathbf{C} \end{pmatrix}$, where $\mathbf{\Phi} = (\mathbf{\Phi}_1, ..., \mathbf{\Phi}_r)$, $\mathbf{\Phi}_i = \mathbf{0}$, $i > p$, $\mathbf{C} = \begin{pmatrix} \mathbf{I}_{k(r-1)} & \mathbf{0}_{k(r-1),k} \end{pmatrix}$. It follows that the transition

equation of the state vector is

$$\boldsymbol{\xi}_t = \boldsymbol{F}\boldsymbol{\xi}_{t-1} + \boldsymbol{e}_t,$$

where $\boldsymbol{e}_t = (\varepsilon_t, \mathbf{0}..., \mathbf{0})'$. Clearly $\boldsymbol{e}_t \sim N(\mathbf{0}, \boldsymbol{\Delta})$, $\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\Omega} & \\ & \mathbf{0}_{k(r-1),k(r-1)} \end{pmatrix}$. On the other hand, the measurement equation has time-varying parameters and dimensions. It takes the form of

$$\mathbf{Z}_t = \boldsymbol{H}_t \boldsymbol{\xi}_t.$$

For $t = q, 2q, 3q, ..., T$, we have $\mathbf{Z}_t = \begin{pmatrix} \mathbf{Y}_{1,t} \\ \mathbf{Y}_{2,t} \end{pmatrix}$, $\mathbf{H}_t = \begin{pmatrix} \mathbf{I}_k & \mathbf{0}_{k,k(q-1)} & \mathbf{0}_{k,k(r-q)} \\ \mathbf{I}_k & \mathbf{D}_{k,k(q-1)} & \mathbf{0}_{k,k(r-q)} \end{pmatrix}$, where $\mathbf{D}_{k,k(q-1)} = (\mathbf{I}_k, ..., \mathbf{I}_k)$.

For $t \neq q, 2q, 3q, ..., T$, we have $\mathbf{Z}_t = \mathbf{Y}_{1,t}$, $\mathbf{H}_t = \begin{pmatrix} \mathbf{I}_k & \mathbf{0}_{k,k(q-1)} & \mathbf{0}_{k,k(r-q)} \end{pmatrix}$.

Then the standard Kalman filter can be used to recursively evaluate the likelihood. The forward recursion consists of the prediction step and update step. The starting point is an assumption on the distribution of the initial state. Assume $\boldsymbol{\xi}_0 \sim N(\boldsymbol{c}_0, \boldsymbol{Q}_0)$. Denote $\boldsymbol{Z}_1^t = (\mathbf{Z}_1', ..., \mathbf{Z}_t')'$. Before the information of Date 1 comes in, the information set $\boldsymbol{Z}_1^0$ is empty, so that $\boldsymbol{\xi}_0 \left| \boldsymbol{Z}_1^0 \right. \sim N\left(\widehat{\boldsymbol{\xi}}_{0|0}, \boldsymbol{P}_{0|0}\right)$, where $\widehat{\boldsymbol{\xi}}_{0|0} = \boldsymbol{c}_0$, $\boldsymbol{P}_{0|0} = \boldsymbol{Q}_0$. At Date t $(t = 1, ..., T)$, we first predict $\boldsymbol{\xi}_t$ and $\mathbf{Z}_t$ conditional on the information set of Date $t-1$.

$$\begin{pmatrix} \boldsymbol{\xi}_t \\ \mathbf{Z}_t \end{pmatrix} = \begin{pmatrix} \boldsymbol{F} \\ \boldsymbol{H}_t \boldsymbol{F} \end{pmatrix} \boldsymbol{\xi}_{t-1} + \begin{pmatrix} \boldsymbol{e}_t \\ \boldsymbol{H}_t \boldsymbol{e}_t \end{pmatrix},$$

It follows that

$$\begin{pmatrix} \boldsymbol{\xi}_t \\ \mathbf{Z}_t \end{pmatrix} \left| \boldsymbol{Z}_1^{t-1} \right. \sim N\left[\begin{pmatrix} \widehat{\boldsymbol{\xi}}_{t|t-1} \\ \widehat{\boldsymbol{Z}}_{t|t-1} \end{pmatrix}, \begin{pmatrix} \boldsymbol{P}_{t|t-1} & \boldsymbol{L}_{t|t-1} \\ \boldsymbol{L}_{t|t-1}' & \boldsymbol{D}_{t|t-1} \end{pmatrix} \right],$$

where

$$\begin{aligned} \widehat{\boldsymbol{\xi}}_{t|t-1} &= \boldsymbol{F}\widehat{\boldsymbol{\xi}}_{t-1|t-1}, \\ \widehat{\boldsymbol{Z}}_{t|t-1} &= \boldsymbol{H}_t \widehat{\boldsymbol{\xi}}_{t|t-1}, \\ \boldsymbol{P}_{t|t-1} &= \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F} + \boldsymbol{\Delta}, \\ \boldsymbol{D}_{t|t-1} &= \boldsymbol{H}_t \boldsymbol{P}_{t|t-1} \boldsymbol{H}_t', \\ \boldsymbol{L}_{t|t-1} &= \boldsymbol{P}_{t|t-1} \boldsymbol{H}_t'. \end{aligned}$$

Then we update $\boldsymbol{\xi}_t$ conditional on $\mathbf{Z}_t$ and $\boldsymbol{Z}_1^{t-1}$. It follows that $\boldsymbol{\xi}_t \left| \boldsymbol{Z}_1^t \right. \sim N\left(\widehat{\boldsymbol{\xi}}_{t|t}, \boldsymbol{P}_{t|t}\right)$, where

$$\begin{aligned}
\widehat{\boldsymbol{\xi}}_{t|t} &= \widehat{\boldsymbol{\xi}}_{t|t-1} + \boldsymbol{L}_{t|t-1}\left(\boldsymbol{D}_{t|t-1}\right)^{-1}\left(\mathbf{Z}_t - \widehat{\boldsymbol{Z}}_{t|t-1}\right), \\
\boldsymbol{P}_{t|t} &= \boldsymbol{P}_{t|t-1} - \boldsymbol{L}_{t|t-1}\left(\boldsymbol{D}_{t|t-1}\right)^{-1}\boldsymbol{L}_{t|t-1}^{'}.
\end{aligned}$$

This completes a recursion cycle and the filter proceeds to the next period. One can also rewrite the recursion formulas in terms of the Kalman gain and Riccati equation by plugging $\widehat{\boldsymbol{\xi}}_{t|t}$ and $\boldsymbol{P}_{t|t}$ back into $\widehat{\boldsymbol{\xi}}_{t+1|t}$ and $\boldsymbol{P}_{t+1|t}$. Once the filter goes through the entire sample periods, we obtain the likelihood function in its prediction error decomposition form, namely $\prod_{t=1}^{T}\phi\left(\mathbf{Z}_t; \widehat{\boldsymbol{Z}}_{t|t-1}, \boldsymbol{D}_{t|t-1}\right)$, where $\phi\left(\boldsymbol{x}; \mu, \boldsymbol{\Sigma}\right)$ is the the density of $N\left(\mu, \boldsymbol{\Sigma}\right)$.

A.4 Simulation studies of varied frequency data

Before we apply the varied frequency data regression on GDP, CPI, federal funds rate and M1 to study the monetary policy shocks, we first generate some pseudo-data to test the performance of our algorithm. We consider a four-variate VAR(1) system, and arbitrarily name the four variates GDP, CPI, federal funds rate and M1 respectively. The data generating process of the reduced form VAR is specified as follows:

$$\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

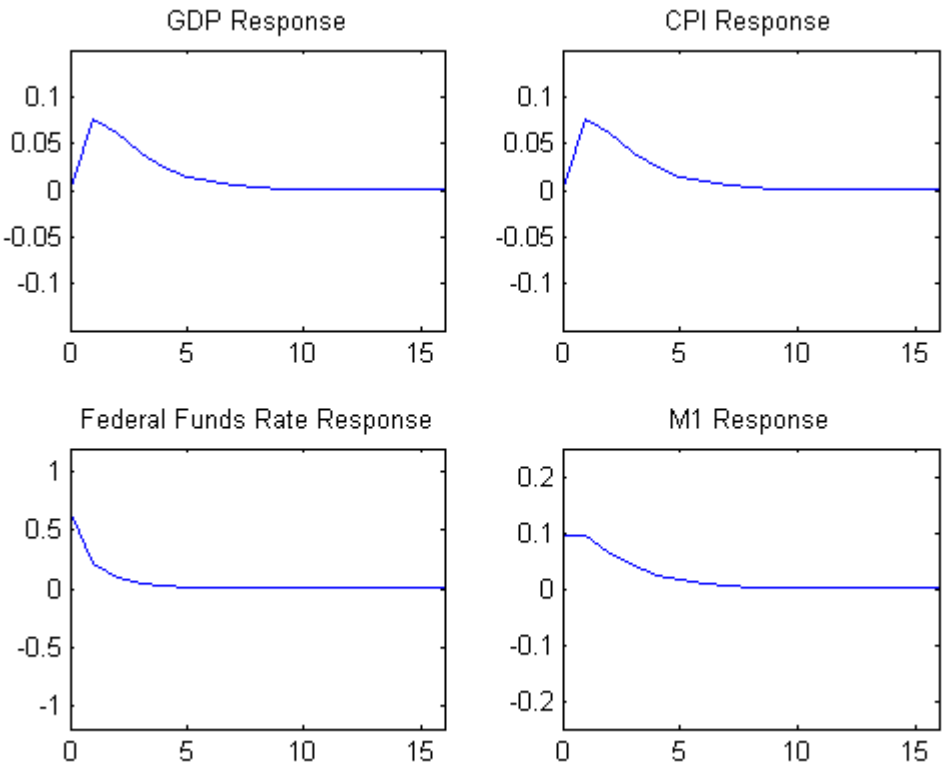
where $\boldsymbol{\Phi} = 0.2 * \boldsymbol{I}_4 + 0.1 * \mathbf{1}_4$, $E\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\right) = 0.8 * \boldsymbol{I}_4 + 0.2 * \mathbf{1}_4$. The symbol $\boldsymbol{I}_4$ refers to a 4×4 identity matrix and $\mathbf{1}_4$ is a 4×4 matrix of ones.

Three thousand observations are generated accordingly. We refer to these observations latent monthly data. To simulate the monthly-quarterly mixture data, we aggregate the first variable (GDP) every three periods, leaving observations of the rest variables (CPI, federal funds rate and M1) intact. To identify the structural model, the recursiveness assumption of zero contemporaneous effects by Christano (1998) is imposed. The order of the four variates implies that the Cholesky decomposition can identify federal funds rate shock. In the simulation study, we already know the true parameters of the VAR model. By inverting the VAR process

into a VMA process, we obtain the theoretical impulse response function. Note that we are not interested in the shape of the dynamic response curve *per se*, since it is just an artificial result of the pre-set parameters. Instead we want to compare the theoretical curve with the estimated impulse response curve using the pseudo-data. Figure A.1 plot the theoretical responses of the four variables to the federal funds shocks.

Then we use the pseudo data to fit two models. The first model is a quarterly data VAR. Despite the availability of the monthly CPI, federal funds rate and M1 data, we aggregate them into quarterly frequency, which is aligned with the GDP data. Then the Bayesian version of the standard VAR model is estimated with these quarterly observations. It is unclear how to set the lag order of the quarterly VAR model, because in theory aggregation introduces moving-average disturbances into the process. We nevertheless use the Akaike and Bayesian information criteria to choose the lag length. Under our data generating process, it seems that both criteria always suggest one lag is the optimal length. So the quarterly VAR(1) specification is adopted. Then the estimated VAR process is inverted into its VMA representation, with the impulse response function plotted in Figure A.2. The solid line represents the sample mean of the response function at each date, and the two dotted lines are the 95% intervals of highest posterior density (HPD).

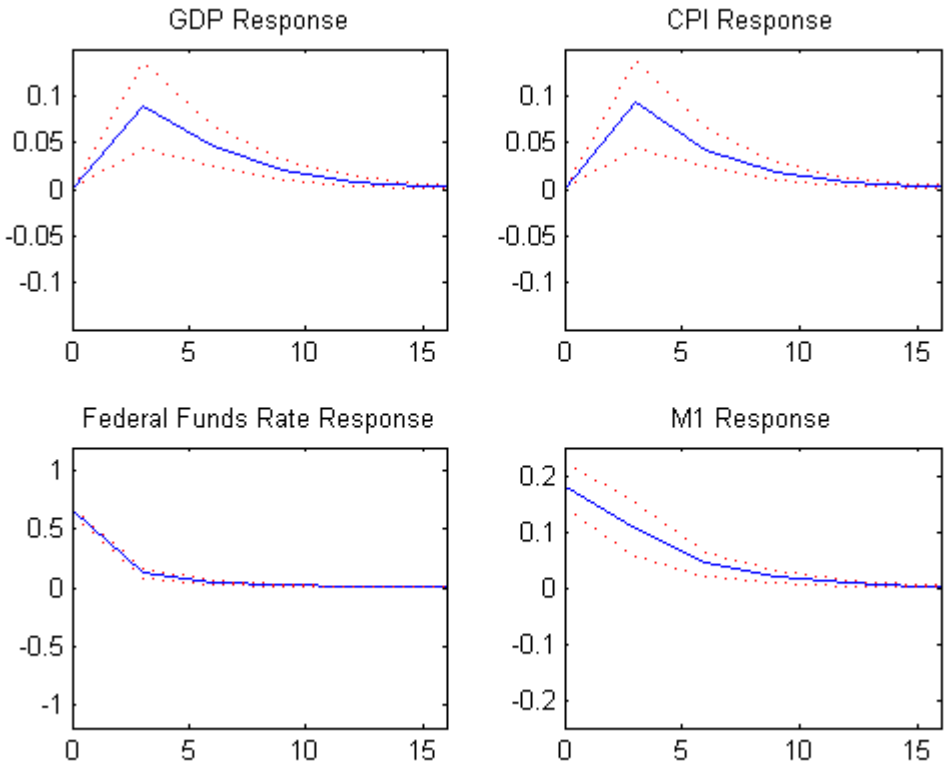
The second model is the varied frequency VAR model that simultaneously uses the quarterly GDP data and the monthly CPI, federal funds rate and M1 data. By our data generating process, we know one lag is the correct choice. Since this is a balanced aggregation, the Gibbs sampler with blocks is employed to simulate the posterior distribution of the model parameters and latent disaggregated variables. After a several thousand MCMC draws, the sample from the posterior conditional distribution begins to stabilize, indicating the convergence of the chain. For accuracy the reported results in Table A.1 and A.2 are based on a hundred thousand draws with the first half burned in. The sample mean of the posterior draws on the autoregressive parameters are reasonably close to the pre-set true values, which always fall within the 95% HPD credible intervals of the posterior distribution of the parameters. The corresponding impulse response functions as well as the 95% HPD intervals are plotted in Figure A.3. Compared with Figure A.2, which is the results under the quarterly VAR model, the



In this simulation study, the reduced form VAR takes the form $\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \varepsilon_t$ with $\boldsymbol{\Phi} = 0.2 * \boldsymbol{I}_4 + 0.1 * \mathbf{1}_4$, $E(\varepsilon_t \varepsilon_t') = 0.8 * \boldsymbol{I}_4 + 0.2 * \mathbf{1}_4$. Under the recursiveness assumption, the structural shocks are identified and the theoretical impulse response functions are then plotted. The names of GDP, CPI, Federal Funds rate and M1 are arbitrary in this exercise.

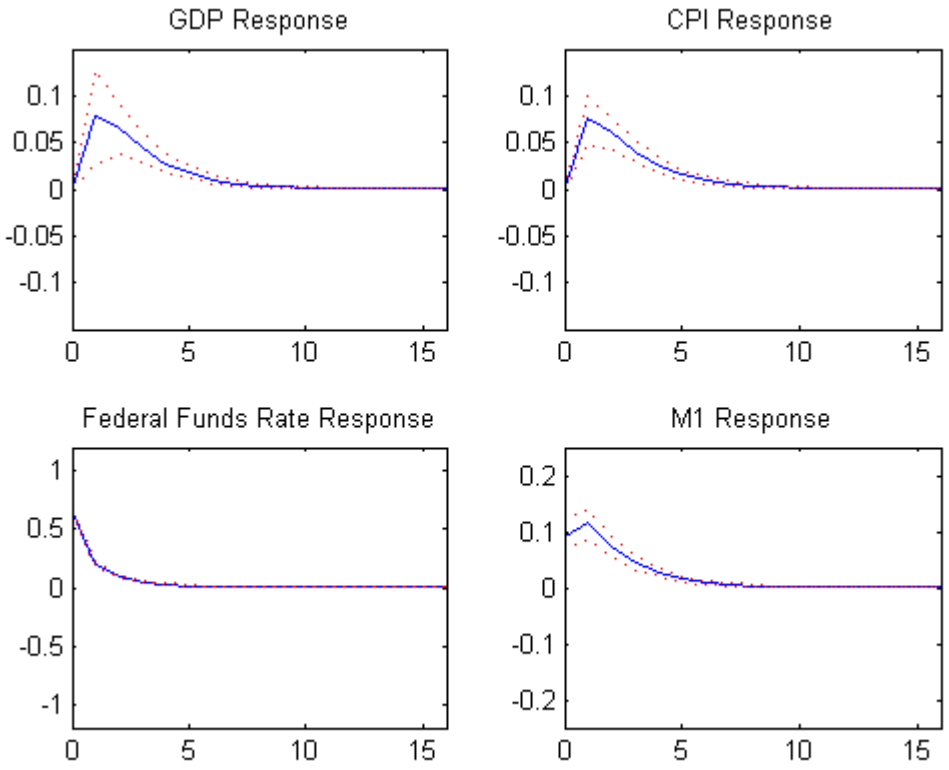
Figure A.1 Theoretical impulse response function in the simulated VAR system

curves in Figure A.3 is clearly closer to the true impulse response curve. Both the shape and the magnitude of the dynamic responses resemble the true curves, indicating the usefulness of the varied frequency VAR approach.



The quarterly VAR system with lags selected by AIC and BIC is fitted by pseudo quarterly data. Under the recursiveness assumption, the structural shocks are identified by the Cholesky decomposition and then the estimated impulse response function is plotted. The names of GDP, CPI, Federal Funds rate and M1 are arbitrary in the simulation exercises. The solid line plots the posterior mean of the impulse-response function and the dotted lines are the 95% HPD credible bands. The results are obtained from a Gibbs sampler of 100000 draws with the first half of draws burned in.

Figure A.2 Dynamic responses to structural shocks with pseudo quarterly data



The varied frequency VAR system is fitted by pseudo monthly-quarterly mixture data. Under the recursiveness assumption, the structural shocks are identified by the Cholesky decomposition and then the estimated impulse response function is plotted. The names of GDP, CPI, Federal Funds rate and M1 are arbitrary in the simulation exercises. The solid line plots the posterior mean of the impulse-response function and the dotted lines are the 95% HPD credible bands. The results are obtained from a Gibbs sampler of 100000 draws with the first half of draws burned in.

Figure A.3 Dynamic responses to structural shocks with pseudo varied frequency data

0.300 0.304 (0.053)	0.100 0.105 (0.037)	0.100 0.084 (0.038)	0.100 0.146 (0.041)
0.100 0.064 (0.035)	0.300 0.313 (0.019)	0.100 0.096 (0.019)	0.100 0.119 (0.020)
0.100 0.096 (0.035)	0.100 0.099 (0.019)	0.300 0.282 (0.018)	0.100 0.129 (0.021)
0.100 0.064 (0.038)	0.100 0.088 (0.020)	0.100 0.127 (0.019)	0.300 0.318 (0.020)

In this simulation exercise, varied frequency VAR is estimated by the Bayesian approach. The true autoregressive coefficients are $\mathbf{\Phi} = 0.2 * \boldsymbol{I}_4 + 0.1 * \mathbf{1}_4$. The above table is divided into 4-by-4 cells. In each cell, the first number is the true parameter value. The second number is the posterior mean of the parameter (Bayesian version of the point estimator) and the third number is the posterior standard deviation of the parameter (Bayesian version of the standard error). The results are obtained from a Gibbs sampler of 100000 draws with the first half of draws burned in.

Table A.1 Autoregressive coefficients estimation using the pseudo varied frequency data

1.000 0.924 (0.084)	0.200 0.175 (0.051)	0.200 0.152 (0.050)	0.200 0.240 (0.055)
0.200 0.175 (0.051)	1.000 0.979 (0.026)	0.200 0.195 (0.018)	0.200 0.175 (0.019)
0.200 0.152 (0.050)	0.200 0.195 (0.018)	1.000 0.970 (0.026)	0.200 0.188 (0.019)
0.200 0.240 (0.055)	0.200 0.175 (0.019)	0.200 0.188 (0.019)	1.000 0.994 (0.026)

In this simulation exercise, varied frequency VAR is estimated by the Bayesian approach. The true covariance matrix is $E\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\right) = 0.8 * \boldsymbol{I}_4 + 0.2 * \mathbf{1}_4$. The above table is divided into 4-by-4 cells. In each cell, the first number is the true parameter value. The second number is the posterior mean of the parameter (Bayesian version of the point estimator) and the third number is the posterior standard deviation of the parameter (Bayesian version of the standard error). The results are obtained from a Gibbs sampler of 100000 draws with the first half of draws burned in.

Table A.2 Covariance matrix estimation using the pseudo varied frequency data

APPENDIX B. ADDITIONAL MATERIAL FOR CHAPTER 3

B.1 Proof of Proposition 3.1

Let $x_i = \mu_i + \varepsilon_i$, where $\varepsilon_i \sim N\left(0, \sigma^2\right)$.

$\overline{x} = \sum_{i=1}^{n} \mu_i + (\varepsilon_1 + ... + \varepsilon_n)$.

Note that $(\varepsilon_1, ..., \varepsilon_n)$ is n dimensional multivariate normal, and so are the n mean-adjusted linear combinations $(x_1, ..., x_{n-1}, \overline{x})$. Then

$$\begin{pmatrix} \mathbf{x}_{-n} \\ \overline{x} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{-n} \\ \sum_{i=1}^{n} \mu_i \end{pmatrix}, \begin{pmatrix} \sigma^2 \mathbf{I}_{n-1} & \sigma^2 \iota_{n-1} \\ \sigma^2 \iota_{n-1}' & n\sigma^2 \end{pmatrix} \right].$$

It follows that

$$\mathbf{x}_{-n} \left| \overline{x} \right. \sim N\left[\mu_{-n} + \frac{1}{n}\left(\overline{x} - \sum_{i=1}^{n} \mu_i \right) \iota_{n-1}, \sigma^2 \left(\mathbf{I}_{n-1} - \frac{1}{n} \iota_{n-1} \iota_{n-1}' \right) \right].$$

Lastly, conditional on $\mathbf{x}_{-n}, \overline{x}$, we have $x_n = \overline{x} - \sum_{i=1}^{n-1} x_i$.

B.2 Proof of Proposition 3.2

Plugging Eq. (3.2) into Eq. (3.1), we have

$$\begin{pmatrix} y_{t,i} \\ x_{t,i} \end{pmatrix} \sim N\left\{ \begin{bmatrix} \mathbf{z}_{t,i}\alpha\beta + \mathbf{w}_{t,i}\left(\beta\gamma + \delta\right) \\ \mathbf{z}_{t,i}\alpha + \mathbf{w}_{t,i}\gamma \end{bmatrix}, \begin{bmatrix} \left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right) & \beta\sigma_v^2 + \sigma_{uv} \\ \beta\sigma_v^2 + \sigma_{uv} & \sigma_v^2 \end{bmatrix} \right\}.$$

It follows that

$$x_{t,i} \left| y_{t,i} \right. \sim N\left(\overline{\mu}_{t,i}, \overline{\sigma}^2 \right),$$

where

$$\overline{\mu}_{t,i} = \mathbf{z}_{t,i}\alpha + \mathbf{w}_{t,i}\gamma + \frac{\beta\sigma_v^2 + \sigma_{uv}}{\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}} \left[y_{t,i} - \mathbf{z}_{t,i}\alpha\beta - \mathbf{w}_{t,i}\left(\beta\gamma + \delta\right) \right],$$

$$\overline{\sigma}^2 = \sigma_v^2 - \left(\beta\sigma_v^2 + \sigma_{uv}\right)^2 \left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}.$$

In the presence of the aggregation constraint, we apply Proposition 3.1. The result follows.

An alternative proof proceeds by deriving the joint distribution of $\left(\mathbf{y}_t', \mathbf{x}_{t,-n}', \overline{x}_t\right)'$, which is a $2n$ dimensional linear combination of $(v_{t,1}, ..., v_{t,n}, u_{t,1}, ..., u_{t,n})$ and thus still multivariate normal. Therefore, the conditional normal distribution of $\mathbf{x}_{t,-n} | \mathbf{y}_t, \overline{x}_t$ can be found after some algebra. The result is the same.

B.3 Comparison of least squares estimators

First, we show that if $\sigma_{uv} \neq 0$, the minimum MSE estimator is inconsistent. To see this, we only need to consider the simplest version of the model. Let $n = 2$; α is a known scalar; regressors $\mathbf{w}_{t,i}$ do not exist. The model becomes:

$$\begin{aligned} y_{t,i} &= x_{t,i}\beta + u_{t,i}, \\ x_{t,i} &= z_{t,i}\alpha + v_{t,i}, \\ \overline{x}_t &= x_{t,1} + x_{t,2}. \end{aligned}$$

For the minimum MSE estimator, the imputed value is

$$\begin{aligned} \widehat{x}_{t,i} &= z_{t,i}\alpha + \frac{1}{2}\left[\overline{x}_t - (z_{t,1}\alpha + z_{t,2}\alpha)\right] \\ &= z_{t,i}\alpha + \frac{1}{2}\left(v_{t,1} + v_{t,2}\right). \end{aligned}$$

In the second step, we regress

$$y_{t,i} = \widehat{x}_{t,i}\beta + \varepsilon_{t,i},$$

where $\varepsilon_{t,i} = u_{t,i} + \beta\left[v_{t,i} - \frac{1}{2}\left(v_{t,1} + v_{t,2}\right)\right]$.

The endogeneity of $\widehat{x}_{t,i}$ to $\varepsilon_{t,i}$ does not come from the presence of $v_{t,1}, v_{t,2}$ in both expressions, but merely from the correlation between $u_{t,i}$ and $v_{t,i}$. To see this, define $\xi_0 = \frac{1}{2}\left(v_{t,1} + v_{t,2}\right)$, and $\xi_i = v_{t,i} - \frac{1}{2}\left(v_{t,1} + v_{t,2}\right)$, $i = 1, 2$. By change of variable, the joint distribution of ξ_0 and ξ_1 is given by

$$\begin{aligned} f\left(\xi_0, \xi_1\right) &= \phi\left(\xi_0 + \xi_1; 0, \sigma_v^2\right) \cdot \phi\left(\xi_0 - \xi_1; 0, \sigma_v^2\right) \cdot |-2| \\ &\propto \exp\left[-\sigma_v^{-2}\left(\xi_0^2 + \xi_1^2\right)\right]. \end{aligned}$$

So ξ_0 and ξ_1 are independent and each distributed as $N\left(0, \frac{1}{2}\sigma_v^2\right)$. Similarly, ξ_0 and ξ_2 are independent.

However, as long as $\sigma_{uv} \neq 0$, we have $cov\left(\widehat{x}_{t,i}, \varepsilon_{t,i}\right) = \frac{1}{2}\sigma_{uv}$, hence the endogenous regressor and inconsistent estimator, no matter whether OLS or GLS is used.

In fact, the OLS version of the estimator

$$\begin{aligned}\widehat{\beta} &= \left(\sum_{t=1}^{T}\sum_{i=1}^{n}\widehat{x}_{t,i}^2\right)^{-1}\left(\sum_{t=1}^{T}\sum_{i=1}^{n}\widehat{x}_{t,i}y_{t,i}\right) \\ &\xrightarrow{p} \beta + \frac{1}{2}\sigma_{uv}\left[\alpha^2 Q_{zz} + \frac{1}{2}\sigma_v^2\right]^{-1},\end{aligned}$$

where $Q_{zz} = E\left(z_{t,i}^2\right)$.

On the other hand, the Dagenais estimator is still consistent. The imputed value is $\widetilde{x}_{t,i} = z_{t,i}\alpha$. Then we regress

$$y_{t,i} = \widetilde{x}_{t,i}\beta + \widetilde{\varepsilon}_{t,i},$$

where $\widetilde{\varepsilon}_{t,i} = u_{t,i} + \beta v_{t,i}$, so that $cov\left(\widetilde{x}_{t,i}, \widetilde{\varepsilon}_{t,i}\right) = 0$, even if $\sigma_{uv} \neq 0$.

The estimator

$$\begin{aligned}\widehat{\beta} &= \left(\sum_{t=1}^{T}\sum_{i=1}^{n}\widetilde{x}_{t,i}^2\right)^{-1}\left(\sum_{t=1}^{T}\sum_{i=1}^{n}\widetilde{x}_{t,i}y_{t,i}\right) \\ &= \beta + \left[\sum_{t=1}^{T}\sum_{i=1}^{n}\left(z_{t,i}\alpha\right)^2\right]^{-1}\cdot\left[\sum_{t=1}^{T}\sum_{i=1}^{n}z_{t,i}\alpha\left(u_{t,i} + \beta v_{t,i}\right)\right],\end{aligned}$$

so that $\widehat{\beta} \xrightarrow{p} \beta$ and

$$\sqrt{nT}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left[0, \left(\alpha^2 Q_{zz}\right)^{-1}\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)\right].$$

Clearly, the asymptotic variance of the Dagenais estimator is increasing with σ_v^2. For large enough σ_v^2, it will exceed the variance of all-aggregated-data estimator, which does not use imputation at all. Therefore, if imputation is of poor quality, there is a possibility that the all-aggregated-data estimator is preferred to the Dagenais estimator.

B.4 Derivation of aggregation of several variables

By marginalization,

$$y_{t,i} = \mathbf{z}_{t,i}\cdot\mathbf{A} + \mathbf{w}_{t,i}\cdot\mathbf{B} + \varepsilon_{t,i}, \varepsilon_{t,i} \sim N\left(0, C\right),$$

where

$$\begin{aligned}
\mathbf{A} &= \sum_{s=1}^{k} \alpha_s \beta_s, \\
\mathbf{B} &= \delta + \sum_{s=1}^{k} \gamma_s \beta_s, \\
C &= \beta' \cdot \mathbf{\Omega} \cdot \beta, \\
\beta &= (1, \beta_1, ..., \beta_k)'.
\end{aligned}$$

Since the disturbance terms are multivariate normal, their (mean adjusted) linear combinations $(y_{t,1}, ..., y_{t,n}, \overline{x}_{1t}, ..., \overline{x}_{kt})$ are also multivariate normal:

$$\begin{pmatrix} \mathbf{y}_t \\ \overline{x}_{1t} \\ ... \\ \overline{x}_{kt} \end{pmatrix} \sim N \left[\begin{pmatrix} \mathbf{z}_{t,i} \cdot \mathbf{A} + \mathbf{w}_{t,i} \cdot \mathbf{B} \\ \overline{\mathbf{z}}_t \cdot \alpha_1 + \overline{\mathbf{w}}_t \cdot \gamma_1 \\ ... \\ \overline{\mathbf{z}}_t \cdot \alpha_k + \overline{\mathbf{w}}_t \cdot \gamma_k \end{pmatrix}, \begin{pmatrix} C \cdot \mathbf{I}_n & \iota_n \cdot (\beta' \cdot \mathbf{\Omega}_{\cdot,-1}) \\ \mathbf{\Omega}'_{\cdot,-1} \cdot \beta \cdot \iota'_n & n \cdot \mathbf{\Omega}_{-1,-1} \end{pmatrix} \right],$$

where $\mathbf{\Omega}_{\cdot,-1}$ is formed by deleting the first column of $\mathbf{\Omega}$, and $\mathbf{\Omega}_{-1,-1}$ is formed by deleting both first row and column of $\mathbf{\Omega}$.

The likelihood can be factorized as

$$f(\mathbf{y}_t, \overline{x}_{1t}, ..., \overline{x}_{kt}) = f(\mathbf{y}_t) \cdot f(\overline{x}_{1t} | \mathbf{y}_t) \cdot ... \cdot f\left(\overline{x}_{kt} \left| \mathbf{y}_t, \overline{x}_{1t}, ..., \overline{x}_{(k-1)t} \right.\right).$$

Clearly, $f(\mathbf{y}_t)$ is a multivariate normal density with the mean $\mathbf{z}_{t,i} \cdot \mathbf{A} + \mathbf{w}_{t,i} \cdot \mathbf{B}$. Using the formula of the conditional normal distribution, we note $f(\overline{x}_{1t} | \mathbf{y}_t)$ is a normal density with the mean being a linear combination of $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$. Similarly, $f(\overline{x}_{2t} | \mathbf{y}_t, \overline{x}_{1t})$ is a normal density with the mean being a linear combination of $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$, $\overline{x}_{1t}$, and $f\left(\overline{x}_{kt} \left| \mathbf{y}_t, \overline{x}_{1t}, ..., \overline{x}_{(k-1)t} \right.\right)$ is also a normal density with the mean being a linear combination of $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$, $\overline{x}_{1t}$,..., $\overline{x}_{(k-1)t}$.

As a result, the analytic ML estimator can be obtained from $k+1$ auxiliary regressions:

Regress $y_{t,i}$ on $\mathbf{z}_{t,i}$, $\mathbf{w}_{t,i}$.

Regress $\overline{x}_{1t}$ on $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$.

Regress $\overline{x}_{2t}$ on $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$, $\overline{x}_{1t}$.

... ...

Regress $\overline{x}_{kt}$ on $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$, $\overline{x}_{1t}$,..., $\overline{x}_{(k-1)t}$.

Suppose $\mathbf{z}_{t,i}$ contains k_z variables, $\mathbf{w}_{t,i}$ contains k_w variables, then the number of reparameterized coefficients are $(k+1)(k_z+k_w)+\frac{k(k+1)}{2}+(k+1)$, which are estimated from auxiliary regressions. By the invariance property of the ML estimator, we can recover $\beta_1, ..., \beta_k$, δ, $\alpha_1, ..., \alpha_k$, $\gamma_1, ..., \gamma_k$, $\mathbf{\Omega}$ as long as $k_z = k$.

APPENDIX C. ADDITIONAL MATERIAL FOR CHAPTER 4

C.1 Proof of Fact 4.1

By the properties of the categorical distribution,

$$E(\mathbf{v}_s) = \mathbf{p}$$

$$Cov(\mathbf{v}_s) = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'$$

Since $\widehat{\mathbf{p}} = \frac{1}{n}\sum_{s=1}^{n} \mathbf{v}_s$, it is a strongly consistent estimator of $\mathbf{p}$, and the central limit theorem implies

$$\sqrt{n}(\widehat{\mathbf{p}} - \mathbf{p}) \xrightarrow{d} N\left[\mathbf{0}, diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'\right].$$

C.2 Proof of Proposition 4.2

The Delta Method implies that

$$\sqrt{n}\left\{\begin{bmatrix} f_L(\widehat{\mathbf{p}}_1) \\ ... \\ f_L(\widehat{\mathbf{p}}_{n_Z}) \end{bmatrix} - \begin{bmatrix} f_L(\mathbf{p}_1) \\ ... \\ f_L(\mathbf{p}_{n_Z}) \end{bmatrix}\right\} \xrightarrow{d} N\left\{\mathbf{0}, \mathbf{G}\left[diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'\right]\mathbf{G}'\right\},$$

where $\mathbf{G}$ is a block diagonal matrix such that

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 & & \\ & \ddots & \\ & & \mathbf{G}_{n_Z} \end{pmatrix}.$$

Since f_L is homogeneous of degree zero, Euler's theorem implies that $\mathbf{G}_i\mathbf{p}_i = 0$, $i = 1, ..., n_Z$.

It follows that $\mathbf{G}\mathbf{p}\mathbf{p}'\mathbf{G}' = \mathbf{0}$. As a result, the covariance matrix simplifies to

$$\mathbf{G}\left[diag\left(\mathbf{p}\right) - \mathbf{p}\mathbf{p}'\right]\mathbf{G}' = \begin{pmatrix} \mathbf{G}_1 \cdot diag\left(\mathbf{p}_1\right) \cdot \mathbf{G}_1' & & \\ & \ddots & \\ & & \mathbf{G}_{n_Z} \cdot diag\left(\mathbf{p}_{n_Z}\right) \cdot \mathbf{G}_{n_Z}' \end{pmatrix}.$$

In the case of the multivariate normal distribution, zero covariance implies independence.

C.3 Proof of Proposition 4.3

From Eq. (4.1),

$$\begin{aligned} f_L\left(\mathbf{p}_i\right) &= E\left(Y \,|Z = z_i, D = d_t\right) \cdot P\left(D = d_t \,|Z = z_i\right) + y_1 \cdot P\left(D \neq d_t \,|Z = z_i\right) \\ &= E\left[Y \cdot I\left(D = d_t\right)|Z = z_i\right] + y_1 \cdot E\left[I\left(D \neq d_t\right)|Z = z_i\right] \\ &= E\left(Q \,|Z = z_i\right) \\ &= \sum_{k=1}^{n_Y}\sum_{m=1}^{n_D} \frac{p_{ikm}}{p_{i\cdot\cdot}} q_{km}. \end{aligned}$$

The last equality is consistent with Eq. (4.2).

Now consider sampling variations. Previously in the paper, we use the encoded vectors $\{\mathbf{v}_s\}_{s=1}^n$ to summarize the sample, which defines $\widehat{\mathbf{p}}$ and $\widehat{\mathbf{p}}_1, ..., \widehat{\mathbf{p}}_{n_Z}$ as well as $f_L\left(\widehat{\mathbf{p}}_i\right)$ accordingly. We can equivalently use i.i.d. $\{Z_s, Y_s, D_s\}_{s=1}^n$ to denote the sample, where the law of (Z_s, Y_s, D_s) is identical to the representative triple (Z, Y, D). Also define

$$Q_s = Y_s \cdot I\left(D_s = d_t\right) + y_1 \cdot I\left(D_s \neq d_t\right).$$

When $\widehat{p}_{i\cdot\cdot} = \frac{1}{n}\sum_{s=1}^n I\left(Z_s = z_i\right) > 0$, the analogue probability estimator $f_L\left(\widehat{\mathbf{p}}_i\right)$ is well-defined and can be written as

$$\begin{aligned} f_L\left(\widehat{\mathbf{p}}_i\right) &= \sum_{k=1}^{n_Y}\sum_{m=1}^{n_D}\left[\frac{\frac{1}{n}\sum_{s=1}^n I\left(Z_s = z_i, Y_s = y_k, D_s = d_m\right)}{\frac{1}{n}\sum_{s=1}^n I\left(Z_s = z_i\right)} q_{km}\right] \\ &= \sum_{k=1}^{n_Y}\sum_{m=1}^{n_D}\left[\frac{\sum_{s=1}^n I\left(Z_s = z_i, Q_s = q_{km}\right)}{\sum_{s=1}^n I\left(Z_s = z_i\right)} q_{km}\right] \\ &= \frac{\sum_{s=1}^n\left[\sum_{k=1}^{n_Y}\sum_{m=1}^{n_D} q_{km} I\left(Q_s = q_{km}\right)\right] \cdot I\left(Z_s = z_i\right)}{\sum_{s=1}^n I\left(Z_s = z_i\right)} \\ &= \frac{\sum_{s=1}^n Q_s \cdot I\left(Z_s = z_i\right)}{\sum_{s=1}^n I\left(Z_s = z_i\right)} \equiv \widetilde{f_L}\left(\mathbf{p}_i\right). \end{aligned}$$

Note that $\widetilde{f_L}(\mathbf{p}_i)$ is simply the analogue moment estimator for $E(Q|Z=z_i)$. It indicates that whether we use analogue probability or analogue moment, the functional form of the estimator is the same. Working on the variance of $\widetilde{f_L}(\mathbf{p}_i)$ is easier than directly computing the variance of $f_L(\widehat{\mathbf{p}}_i)$.

To make notations compact, denote $\theta \equiv f_L(\mathbf{p}_i)$, $\widetilde{\theta} \equiv \widetilde{f_L}(\mathbf{p}_i) = f_L(\widehat{\mathbf{p}}_i)$, $\gamma \equiv Var(Q|Z=z_i)$.

From here to the end of the proof, when we write $E(\cdot)$, we leave implicit that the expectation is conditional on $\widehat{p}_{i\cdot\cdot} > 0$.

Using the law of iterated expectations, we have

$$\begin{aligned} E\left(\widetilde{\theta}\right) &= E\left[E\left(\widetilde{\theta}\,|\{Z_s\}_{s=1}^n\right)\right] \\ &= E\left[\frac{\sum_{s=1}^n \theta I(Z_s=z_i)}{\sum_{s=1}^n I(Z_s=z_i)}\right] \\ &= \theta. \end{aligned}$$

Then the variance of $\widetilde{\theta}$ equals

$$\begin{aligned} Var\left(\widetilde{\theta}\right) &= E\left[E\left(\widetilde{\theta}^2\,|\{Z_s\}_{s=1}^n\right)\right] - \theta^2 \\ &= E\left\{\frac{\sum_{a=1}^n\sum_{b=1}^n E(Q_aQ_b|\{Z_s\}_{s=1}^n) I(Z_a=z_i) I(Z_b=z_i)}{\sum_{a=1}^n\sum_{b=1}^n I(Z_a=z_i) I(Z_b=z_i)}\right\} - \theta^2 \\ &= E\left\{\frac{\sum_{a=1}^n\sum_{b=1}^n \theta^2 I(Z_a=z_i) I(Z_b=z_i) + \sum_{a=1}^n \gamma I(Z_a=z_i)}{\sum_{a=1}^n\sum_{b=1}^n I(Z_a=z_i) I(Z_b=z_i)}\right\} - \theta^2 \\ &= E\left[\frac{1}{\sum_{a=1}^n I(Z_a=z_i)}\right]\cdot\gamma \\ &= \left[\sum_{r=1}^n \frac{1}{r}\frac{\binom{n}{r}(p_{i\cdot\cdot})^r(1-p_{i\cdot\cdot})^{n-r}}{1-(1-p_{i\cdot\cdot})^n}\right]\cdot\gamma \end{aligned}$$

Note that in the second and third equality, $E(Q_aQ_b|\{Z_s\}_{s=1}^n)$ itself does not equal to $E(Q_aQ_b|Z_a=z_i,Z_b=z_i)$. However, $E(Q_aQ_b|\{Z_s\}_{s=1}^n) I(Z_a=z_i) I(Z_b=z_i)$ equals $E(Q_aQ_b|Z_a=z_i,Z_b=$ For $a\neq b$, $E(Q_aQ_b|Z_a=z_i,Z_b=z_i)=\theta^2$; for $a=b$, $E(Q_aQ_b|Z_a=z_i,Z_b=z_i)=\theta^2+\gamma$. The results follows.

C.4 Proof of Proposition 4.4

Jensen's inequality implies $B_1(\boldsymbol{\mu})$ is bounded below by zero. To show it is also bounded above, we first show $E[T_1(\mathbf{X})]$ is strictly increasing in each μ_i. As the maximum of j normal

variates, $T_1(\mathbf{X})$ has the c.d.f.

$$F(c;\boldsymbol{\mu}) = \prod_{i=1}^{j} P(X_i \leq c) = \prod_{i=1}^{j} \Phi\left(c-\mu_i; 0, \sigma_i^2\right).$$

Since the normal c.d.f. is a strictly increasing function, $F(c;\boldsymbol{\mu})$ is strictly decreasing in $\boldsymbol{\mu}$. To evaluate the expectation, we use the formula, as is suggested by David (1981) and Ross (2010),

$$E[T_1(\mathbf{X})] = \int_0^{\infty} [1 - F(c;\boldsymbol{\mu}) - F(-c;\boldsymbol{\mu})]\, dc,$$

It follows that $E[T_1(\mathbf{X})]$ is strictly increasing in each μ_i. Also note that $\max(\boldsymbol{\mu})$ is merely non-decreasing in each μ_i. Therefore, to maximize $B_1(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$, a necessary condition is $\mu_a = \mu_b$, $\forall a, b = 1, ..., j$. Otherwise, consider $\mu_a < \mu_b$, for some a, b. Let $\Delta' = \mu_b - \mu_a$, then increasing μ_a by Δ' will increase $E[T_1(\mathbf{X})]$ while leaving $\max(\boldsymbol{\mu})$ unchanged. A contradiction to the maximum.

Lastly, by the property of the $\max(\cdot)$ function,

$$\begin{aligned} B_1(\boldsymbol{\mu}+c\cdot\boldsymbol{\iota}) &= E[T_1(\mathbf{X}) + c] - [\max(\boldsymbol{\mu}) + c] \\ &= B_1(\boldsymbol{\mu}), \end{aligned}$$

$\forall c \in \mathbb{R}$, where $\boldsymbol{\iota}$ is a vector of ones. This implies as long as $\mu_a = \mu_b \equiv \mu_0$, $\forall a, b = 1, ..., j$, $B_1(\cdot)$ does not depend on specific choice of μ_0. We pick $\mu_0 = 0$, and $B_1(\mathbf{0})$ attains the maximum $E[\max(\mathbf{X}_0)]$.

C.5 Proof of Fact 4.5

$$\begin{aligned} B_2(\boldsymbol{\mu}) &= E[T_1(\mathbf{X}) - B_1(\mathbf{X})] - \max(\boldsymbol{\mu}) \\ &= B_1(\boldsymbol{\mu}) - E[B_1(\mathbf{X})]. \end{aligned}$$

Proposition **??** indicates that $B_1(\boldsymbol{\mu}) > 0$, $\forall \boldsymbol{\mu} \in \mathbb{R}^j$, so that $E[B_1(\mathbf{X})] > 0$. So we have $B_2(\boldsymbol{\mu}) < B_1(\boldsymbol{\mu})$.

C.6 Proof of Proposition 4.6

To show the proposition, we first introduce a lemma.

Lemma: The n^{th} (uncentered) moment of $N\left(\mu,\sigma^2\right)$ is a polynomial of order n with respect to μ. The leading coefficient (that of μ^n) is one, and the next leading coefficient (that of μ^{n-1}) is zero.

Proof: It is well known that the central moment of $N\left(\mu,\sigma^2\right)$ has a closed-form expression.

$$E\left[(X-\mu)^n\right]=\begin{cases}0 & \text{if } n \text{ is odd}\\ \sigma^d\,(n-1)!! & \text{if } n \text{ is even}\end{cases},$$

where $(n-1)!!$ is the double factorial. This implies that $E\left[(X-\mu)^n\right]$ is a constant with respect to μ. To find the raw moment $E\left(X^n\right)$, we expand $E\left[(X-\mu)^n\right]$ with the formula

$$(a+b)^n=\sum_{k=0}^{n}\binom{n}{k}a^{n-k}b^k.$$

Put $a=1$, $b=-1$, we have $\sum_{k=0}^{n}\binom{n}{k}(-1)^k=0$, or $\sum_{k=1}^{n}\binom{n}{k}(-1)^k=-1$. We will show the lemma by induction. Clearly, the it holds for $n=1$. Suppose it is true for the first $n-1$ raw moments, we want to show it holds for the n^{th} raw moment. Note that

$$E\left[(X-\mu)^n\right]=E\left(X^n\right)+\sum_{k=1}^{n}\binom{n}{k}(-\mu)^k\,E\left(X^{n-k}\right).$$

As is assumed, $E\left(X^{n-k}\right)$ is a polynomial of order $n-k$, the leading coefficient is one and the next leading coefficient is zero, hence $\sum_{k=1}^{n}\binom{n}{k}(-\mu)^k\,E\left(X^{n-k}\right)$ is a polynomial of order n, the leading coefficient is $\sum_{k=1}^{n}\binom{n}{k}(-1)^k=-1$, and the next leading coefficient is zero. It follows that $E\left(X^n\right)$ is a polynomial of order n, with the leading coefficient being one and the next leading coefficient being zero. This proves the lemma.

Now put $r=2$ and consider $B_r\left(\boldsymbol{\mu}\right)=B_{r-1}\left(\boldsymbol{\mu}\right)-E\left[B_{r-1}\left(\mathbf{X}\right)\right]$. Since $B_{r-1}\left(\boldsymbol{\mu}\right)$ is a polynomial of order d w.r.t. $\boldsymbol{\mu}$, so the leading term takes the form $\prod_{i=1}^{j}\mu_i^{a_i}$, where $\sum_{i=1}^{j}a_i=d$. The corresponding term in $E\left[B_{r-1}\left(\mathbf{X}\right)\right]$ takes the form $E\left(\prod_{i=1}^{j}X_i^{a_i}\right)=\prod_{i=1}^{j}E\left(X_i^{a_i}\right)$. By the lemma, $E\left(X_i^{a_i}\right)$ is a polynomial of order a_i w.r.t. μ_i, and the coefficient of the leading term $\mu_i^{a_i}$ is

one, and the coefficient of the next leading term $\mu_i^{a_i-1}$ is zero. This implies that $\prod_{i=1}^{j} E\left(X_i^{a_i}\right)$ is a polynomial of order d w.r.t. $\boldsymbol{\mu}$, with the leading term (of order d) coefficient one and next leading terms (of order $d-1$) zero. As a result, when $B_{r-1}(\boldsymbol{\mu})$ is subtracted by $E\left[B_{r-1}(\mathbf{X})\right]$, the terms corresponding to order d and $d-1$ are canceled, so the order of the polynomial is reduced by 2. The same arguments can be applied to $r = 3, 4, 5$, etc.

C.7 Proof of Proposition 4.7

Let $A \equiv E\left[g\left(\xi_i, \eta_i\right)\right] = E\left[g\left(\xi_j, \eta_{j,k}\right)\right]$, $B \equiv Var\left[g\left(\xi_i, \eta_i\right)\right] = Var\left[g\left(\xi_j, \eta_{j,k}\right)\right]$, $\forall i = 1, ..., N^2$, $j = 1, ..., N$, $k = 1, ..., N$. The two equalities hold because $\left(\xi_j, \eta_{j,k}\right)$ are drawn by the method of composition, the joint distribution of $\left(\xi_j, \eta_{j,k}\right)$ is the same as that of directly sampled $\left(\xi_i, \eta_i\right)$. Clearly, $E\left(S_1\right) = E\left(S_2\right) = A$, $Var\left(S_1\right) = \frac{1}{N^2}B$. When we compute $Var\left(S_2\right)$, we need to consider the covariance terms as well.

$$\begin{aligned} Var\left(S_2\right) &= \frac{1}{N} Var\left[\frac{1}{N}\sum_{k=1}^{N} g\left(\xi_1, \eta_{1,k}\right)\right] \\ &= \frac{1}{N}\frac{1}{N^2}\sum_{k=1}^{N}\sum_{h=1}^{N} cov\left[g\left(\xi_1, \eta_{1,k}\right), g\left(\xi_1, \eta_{1,h}\right)\right] \\ &= \frac{1}{N^2}B + \frac{1}{N^3}\sum_{k=1}^{N}\sum_{h=1, h\neq k}^{N} cov\left[g\left(\xi_1, \eta_{1,k}\right), g\left(\xi_1, \eta_{1,h}\right)\right]. \end{aligned}$$

Now we show each of those covariance terms is non-negative.

$$\begin{aligned} & cov\left[g\left(\xi_1, \eta_{1,k}\right), g\left(\xi_1, \eta_{1,h}\right)\right] \\ &= E\left\{\left[g\left(\xi_1, \eta_{1,k}\right) - A\right] \cdot \left[g\left(\xi_1, \eta_{1,h}\right) - A\right]\right\} \\ &= E_{\xi_1}\left\{E_{\eta_{1,k}|\xi_1}\left[g\left(\xi_1, \eta_{1,k}\right) - A\right] \cdot E_{\eta_{1,h}|\xi_1}\left[g\left(\xi_1, \eta_{1,h}\right) - A\right]\right\} \\ &= E_{\xi_1}\left[c^2\left(\xi_1\right)\right] \geq 0, \end{aligned}$$

where $c\left(\xi_1\right) \equiv E_{\eta_{1,k}|\xi_1}\left[g\left(\xi_1, \eta_{1,k}\right) - A\right] = E_{\eta_{1,h}|\xi_1}\left[g\left(\xi_1, \eta_{1,h}\right) - A\right]$. It follows that $Var\left(S_1\right) \leq Var\left(S_2\right)$.

Note that in the above proof, $Var\left(S_1\right) = Var\left(S_2\right)$ only if $E_{\eta_{1,k}|\xi_1}\left[g\left(\xi_1, \eta_{1,k}\right)\right] = A$ for all realizations of ξ_1. The independency of ξ and η does not necessarily imply $Var\left(S_1\right) = Var\left(S_2\right)$.

When we take conditional expectation of $g(\xi_1, \eta_{1,k})$, ξ_1 should be treated as a constant and in general $c(\xi_1) \neq 0$, even if for independent variates.

BIBLIOGRAPHY

Allison, P. D., 2000. Multiple imputation for missing data: a cautionary tale. Sociological methods and Research 28, 301–309.

Amemiya, T., Wu, R. Y., 1972. The effect of aggregation on prediction in the autoregressive model. Journal of the American Statistical Association 67 (339), 628–632.

Anderson, T. W., 1957. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. Journal of the American Statistical Association 52 (278), 200–203.

Andreou, E., Ghysels, E., Kourtellos, A., 2010. Regression models with mixed sampling frequencies. Journal of Econometrics 158 (2), 246–261.

Aruoba, S. B.and Diebold, F. X., Scotti, C., 2009. Real-time measurement of business conditions. Journal of Business & Economic Statistics 27 (4), 417–427.

Breitung, J., Swanson, N., 2002. Temporal aggregation and spurious instantaneous causality in multiple time series models. Journal of Time Series Analysis 23 (6), 651–665.

Bureau of Labor Statistics, U.S. Department of Labor, 2010. Occupational Outlook Handbook. Washington: U.S. Government Printing Office.

Cain, M., 1994. The moment-generating function of the minimum of bivariate normal random variables. The American Statistician 48, 124–125.

Camacho, M., Perez-Quiros, G., 2010. Introducing the euro-sting: Short-term indicator of euro area growth. Journal of Applied Econometrics 25 (4), 663–694.

Chen, B., Zadrozny, P., 1998. An extended yule-walker method for estimating a vector autoregressive model with mixed-frequency data. In: NBER/NSF Time Series Conference.

Chernozhukov, V., Lee, S. S., Rosen, A., 2009. Intersection bounds: estimation and inference. CeMMAP working papers (CWP19/09).

Chiu, C., Eraker, B., Foerster, A., Kim, T. B., Seoane, H., 2011. Estimating var sampled at mixed or irregular spaced frequencies: A bayesian approach (manuscript).

Christiano, L. J., Eichenbaum, M., Evans, C. L., 1998. Monetary policy shocks: What have we learned and to what end? NBER Working Papers (6400).

Clark, C. E., 1961. The greatest of a finite set of random variables. Operations Research 9, 145–162.

Clements, M. P., Galvao, A. B., 2008. Macroeconomic forecasting with mixed-frequency data. Journal of Business and Economic Statistics 26, 546–554.

Clements, M. P., Galvao, A. B., 2009. Forecasting us output growth using leading indicators: an appraisal using midas models. Journal of Applied Econometrics 24 (7), 1187–1206.

Dagenais, M. G., 1973. The use of incomplete observations in multiple regression analysis : A generalized least squares approach. Journal of Econometrics 1 (4), 317–328.

David, H. A., 1981. Order Statistics. John Wiley & Sons: Hoboken.

Davidson, R., MacKinnon, J., 2002. Fast double bootstrap tests of nonnested linear regression models. Econometric Reviews 21 (4), 419–429.

Davidson, R., MacKinnon, J. G., 2007. Improving the reliability of bootstrap tests with the fast double bootstrap. Computational Statistics & Data Analysis 51 (7), 3259–3281.

Eichenbaum, M., 1992. Comment on interpreting the macroeconomic time series facts : The effects of monetary policy. European Economic Review 36 (5), 1001–1011.

Fraser, D. A. S., 1951. Normal samples with linear constraints and given variances. Canadian Journal of Mathematics 3, 363–366.

Geweke, J. F., 1978. Temporal aggregation in the multiple regression model. Econometrica 46 (3), 643–61.

Geweke, J. F., 1995. Bayesian inference for linear models subject to linear inequality constraints. Working Papers 552, Federal Reserve Bank of Minneapolis.

Ghysels, E., Santa-Clara, P., Valkanov, R., 2006. Predicting volatility: getting the most out of return data sampled at different frequencies. Journal of Econometrics 131 (1-2), 59–95.

Ghysels, E., Sinko, A., Valkanov, R., 2007. Midas regressions: Further results and new directions. Econometric Reviews 26 (1), 53–90.

Gomez, V., Maravall, A., Pena, D., 1998. Missing observations in arima models: Skipping approach versus additive outlier approach. Journal of Econometrics 88 (2), 341–363.

Gourieroux, C., Monfort, A., 1981. On the problem of missing data in linear models. Review of Economic Studies 48 (4), 579–86.

Hamilton, J. D., 1994. Time Series Analysis. Princeton University Press: Princeton.

Harvey, A. C., Pierse, R. G., 1984. Estimating missing observations in economic time series. Journal of the American Statistical Association 79 (385), 125–131.

Hsiao, C., 1979. Linear regression using both temporally aggregated and temporally disaggregated data. Journal of Econometrics 10 (2), 243–252.

Hyung, N., Granger, C. W., 2008. Linking series generated at different frequencies. Journal of Forecasting 27 (2), 95–108.

Jones, R. H., 1980. Maximum likelihood fitting of arma models to time series with missing observations. Technometrics 22 (3), 389–395.

Koop, G., Poirier, D. J., Tobias, J. L., 2007. Bayesian Econometric Methods. Cambridge Books. Cambridge University Press.

Kreider, B., Pepper, J. V., 2007. Disability and employment: Reevaluating the evidence in light of reporting errors. Journal of the American Statistical Association 102, 432–441.

Kuzin, V., Marcellino, M., Schumacher, C., 2011. Midas vs. mixed-frequency var: Nowcasting gdp in the euro area. International Journal of Forecasting 27 (2), 529–542.

Manski, C. F., 1997. Monotone treatment response. Econometrica 65 (6), 1311–1334.

Manski, C. F., Pepper, J. V., 2000. Monotone instrumental variables, with an application to the returns to schooling. Econometrica 68 (4), 997–1012.

Manski, C. F., Pepper, J. V., 2009. More on monotone instrumental variables. Econometrics Journal 12 (s1), 200–216.

Marcellino, M., 1999. Some consequences of temporal aggregation in empirical analysis. Journal of Business & Economic Statistics 17 (1), 129–136.

Marcellino, M., Schumacher, C., 2010. Factor midas for nowcasting and forecasting with ragged-edge data: A model comparison for german gdp. Oxford Bulletin of Economics and Statistics 72 (4), 518–550.

Mariano, R. S., Murasawa, Y., 2003. A new coincident index of business cycles based on monthly and quarterly series. Journal of Applied Econometrics 18 (4), 427–443.

Mariano, R. S., Murasawa, Y., 2010. A coincident index, common factors, and monthly real gdp. Oxford Bulletin of Economics and Statistics 72 (1), 27–46.

Mittnik, S., Zadrozny, P. A., 2004. Forecasting quarterly german gdp at monthly intervals using monthly ifo business conditions data. CESifo Working Paper Series.

Palm, F. C., Nijman, T. E., 1982. Linear regression using both temporally aggregated and temporally disaggregated data. Journal of Econometrics 19 (2-3), 333–343.

Pavia-Miralles, J., 2010. A survey of methods to interpolate, distribute and extrapolate time series. Journal of Service Science and Management 3, 449–463.

Pena, D., Tiao, G. C., 1991. A note on likelihood estimation of missing values in time series. Archivo Institucional de la Universidad Carlos III de Madrid, Working Papers, 1991-01.

Proietti, T., 2008. Missing data in time series: A note on the equivalence of the dummy variable and the skipping approaches. Statistics and Probability Letters 78 (3), 257–264.

Ross, A., 2010. Computing bounds on the expected maximum of correlated normal variables. Methodology and Computing in Applied Probability 12, 111–138.

Rubin, D. B., 1987. Multiple imputation for nonresponse in surveys. New York: Wiley.

Sargan, J. D., Drettakis, E. G., 1974. Missing data in an autoregressive model. International Economic Review 15 (1), 39–58.

Schafer, J. L., 1997. Analysis of incomplete multivariate data. Chapman and Hall: London.

Silvestrini, A., Veredas, D., 2008. Temporal aggregation of univariate and multivariate time series models: A survey. Journal of Economic Surveys 22 (3), 458–497.

Sims, C. A., 1980. Macroeconomics and reality. Econometrica 48 (1), 1–48.

Sims, C. A., 1992. Interpreting the macroeconomic time series facts: The effects of monetary policy. European Economic Review 36 (5), 975–1000.

Stoica, P., Xu, L., Li, J., 2005. A new type of parameter estimation algorithm for missing data problems. Statistics & Probability Letters 75 (3), 219–229.

Tiao, G. C., Wei, W. S., 1976. Effect of temporal aggregation on the dynamic relationship of two time series variables. Biometrika 63 (3), 513–523.

Van Buuren, S., Boshuizen, H. C., Knook, D. L., 1999. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 18, 681–694.

Viefers, P., 2011. Bayesian inference for the mixed-frequency var model. Discussion Papers of DIW Berlin 1172.

Zadrozny, P., 1988. Gaussian likelihood of continuous-time armax models when data are stocks and flows at different frequencies. Econometric Theory 4 (01), 108–124.